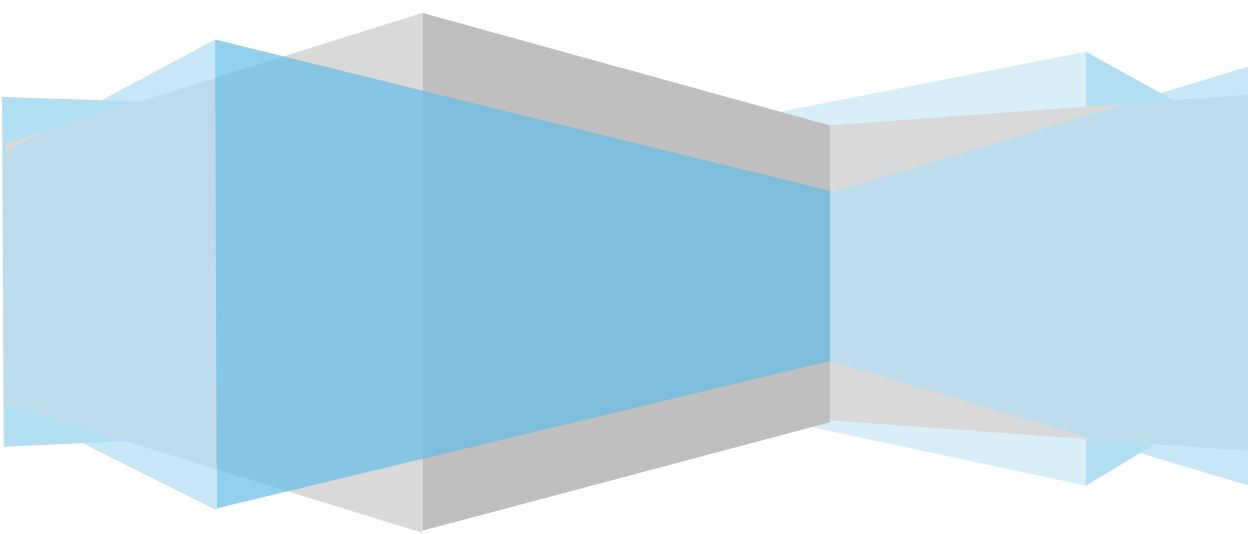
MATHEMATICAL MODELING IN PSYCHOLOGY USING ARTIFICIAL INTELLIGENCE

AGNIESZKA SZYMAŃSKA

UKSW

Mathematical Modeling in Psychology Using Artificial Intelligence



Cardinal Stefan Wyszynski University in Warsaw

Mathematical Modeling in Psychology Using Artificial Intelligence

Agnieszka Szymańska

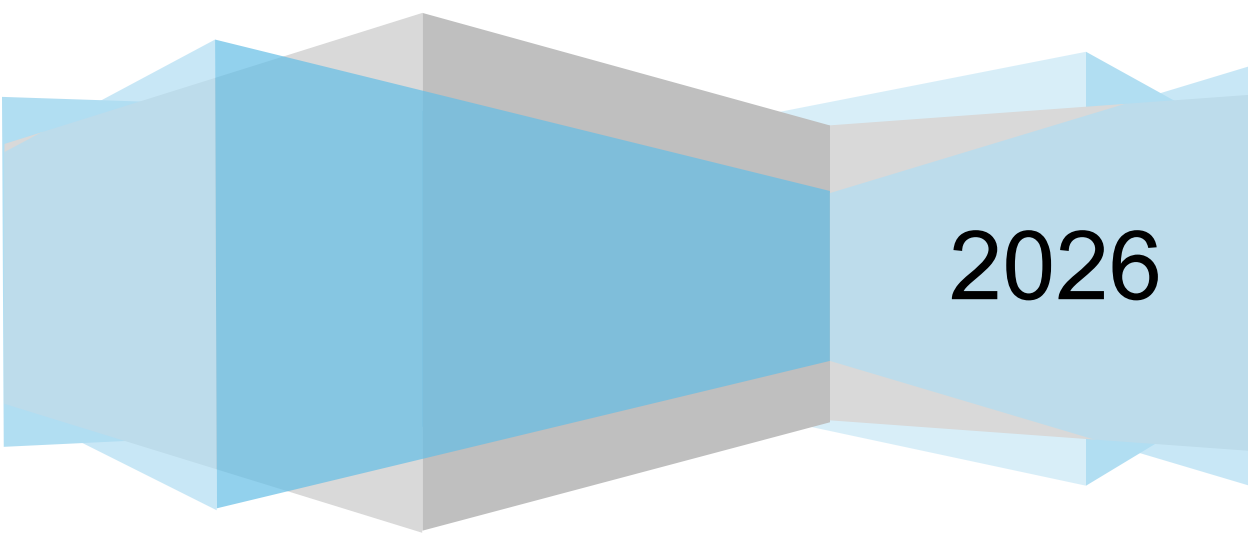


TABLE OF CONTENTS

PREFACE

This book on mathematical modeling in psychology using artificial intelligence aims to demonstrate how AI algorithms can support the process of model construction and analysis in psychology and psychometrics. The inspiration for its creation came from my fifteen years of experience working with artificial intelligence algorithms and mathematical modeling methods in psychology. Over the years, I have not only explored the theoretical and practical aspects of these algorithms but also passed on this knowledge during numerous teaching activities.

My experience includes teaching at both the master's and doctoral levels at the Cardinal Stefan Wyszyński University in Warsaw (UKSW) – in the Institute of Psychology as well as within the Big Data in Social Analytics program. During these courses, conducted in both Polish and English, including within the Erasmus program, I had the opportunity not only to introduce students to the functioning and applications of artificial intelligence algorithms but also to demonstrate how these tools can be practically applied in psychological research. It was from these academic and research experiences that the need arose to create a book that could serve as a guide for those who wish to understand and implement modern data analysis techniques in psychology.

In addition to my teaching work, over the years I have also had many opportunities to verify and test solutions involving artificial intelligence algorithms and other mathematical algorithms across various fields of psychology. Extensive collaboration with outstanding researchers in Poland and abroad, involved in research projects, enabled me to gain a deeper understanding of these issues and to apply the acquired methods in practice.

From a methodological perspective, collaboration with Professor Elżbieta Aranowska and Dr Rytel proved particularly valuable, resulting in an in-depth analysis and application of artificial intelligence algorithms and other advanced measurement methods, which are described in detail in this book.

At the same time, working with Professor Lidia Grzesiuk in the field of psychotherapy opened new perspectives for the application of these algorithms, particularly in the context of analysing psychotherapeutic processes and research on workplace mobbing. Collaboration with Professor Barbara Bokus in the area of psycholinguistics, in turn, enabled the use of artificial intelligence algorithms—such as decision trees and text mining techniques—for the analysis of language and cognitive processes.

Moreover, my own research in upbringing psychology—particularly on parental mistakes and parental goals—led me to apply the theory developed by Professor Antonina Gurycka, which I tested using artificial intelligence algorithms. These studies allowed for a better understanding and modeling of complex relationships within upbringing processes and represented a significant contribution to the development of this field, as confirmed by my PhD and Doctor of Science (DSc, habilitation) degrees.

All of these experiences have not only enriched my understanding of artificial intelligence in psychology, but have also strengthened my conviction that integrating these modern tools with traditional psychological methods is not only possible, but essential for the continued advancement of science.

As a psychologist and a graduate of postgraduate studies in mathematics, I have always appreciated the value of combining and integrating psychology with mathematics—especially for the purpose of gaining deeper insight into psychological dependencies that had previously remained undescribed. In both my research and teaching practice, I observed that artificial intelligence algorithms integrate remarkably well with a wide range of statistical methods and, moreover, can significantly enhance and refine the analytical process.

This approach enabled me to expand certain models that had not previously been widely applied. Together with other professors, as well as individually, I developed several original analytical methods that will be presented in this book. One such example is the use of grade correspondence analysis to construct structural models—that is, structural equation systems—which can be derived at the exploratory level. This is my own original idea, in which I demonstrate how a mathematical algorithm can support the process of creating a structural model by enabling its development at the exploratory, rather than theoretical, stage.

The creation of exploratory models is extremely time-consuming and requires a high degree of precision. The grade algorithm can be used effectively to build an exploratory model that significantly reduces the time needed to construct it, while simultaneously ensuring optimal links between variables. This algorithm makes it possible to quickly and accurately identify the most important relationships within a dense network of interdependencies, thereby making the process of discovering and validating models more efficient and productive.

This book offers a detailed discussion of how these and other original methods can be applied in psychological research, and how combining traditional approaches with modern artificial intelligence algorithms opens new perspectives in psychology and psychometrics.

Another example, which will also be discussed in detail, concerns the solution to the problem of matching model curves to empirical curves. Formulas for fitting model curves to theoretical ones, developed by Professor Elżbieta Aranowska, have existed since the 1980s (Aranowska, 1989). However, the difficulty lay in the fact that deriving empirical curves was extremely challenging and time-consuming, primarily due to the need to select an appropriate construction method and to manage variance and data scatter.

The emergence of new analytical algorithms, including clustering methods such as *k*-means, opened up new possibilities in this area. In particular, an original idea involving the application of the *k*-means algorithm and the expectation–maximisation (EM) algorithm to derive empirical curves proved crucial. Of special importance was the use of the normalised mean, which made it possible to convert results into a common scale, thereby creating a foundation for profiling and for the subsequent use of data in analysing the fit between model curves and empirical ones. Although *k*-means is not, in the strict sense, an artificial intelligence algorithm, its application in combination with modern data exploration techniques significantly contributed to the development of procedures for fitting model curves to empirical ones (Szymańska, 2023a).

As a result, it became possible to profile empirical curves that faithfully reflect real phenomena observed in the population. This innovative approach to using exploratory tools for curve fitting will be described in detail in this book, demonstrating how modern methods can revolutionise classical approaches in psychometrics and mathematical modeling in psychology, including applications in profiling within criminology. Other innovative topics will also be discussed, including a previously unpublished solution by Professor Elżbieta Aranowska concerning the determination of path validity in a structural model, as well as a novel method for verifying circular models using grade algorithms and support vector machines.

These topics, combined with classical discussions of artificial intelligence algorithms and general mathematical modeling—including structural equation systems—form the core of this book. We will begin by discussing the formal assumptions of structural models and their limitations. The second part will focus on inductive algorithms such as decision trees. Next, we will turn to classification algorithms, including cluster analyses, grade algorithms, and other methods of classification and clustering. The fourth part of the book will be devoted to machine learning methods: neural networks, support vector machines, and other predictive techniques.

Subsequently, we will discuss speech recognition algorithms, including language models such as GPT and BERT, with particular emphasis on their applications and differences. This section will also address text mining and the transformation of qualitative data into quantitative form.

The conclusion of the book will focus on network analytics in the context of Social Network Analysis (SNA). We will show how SNA is constructed and applied in data analysis, as well as its potential applications in psychology. The final chapter will concentrate on psychometric issues, where I will share my experience in applying artificial intelligence algorithms in contemporary psychometrics.

It should be noted that psychology and artificial intelligence constitute a relatively new area of convergence. Most of the examples and solutions presented in this book are based on my own research and experience, which makes this publication a pioneering work. In this context, it is worth noting the work of researchers such as Hoi Yin Bonnie Yim et al. (2014) and Liu and Wei (2023), who emphasise the importance of such innovative research for the development of psychology.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor Elżbieta Aranowska, my doctoral supervisor, for her many years of collaboration and for introducing me to the world of science. I thank her with all my heart for her guidance, mentorship, and support, which were invaluable throughout my academic journey. As an outstanding mathematician, statistician, psychometrician, and methodologist, Professor Aranowska had a profound impact on the development of my knowledge and skills. Our many years of cooperation resulted in the development of numerous methods, some of which are described in this book. The idea for this book also emerged from our shared discussions. Sadly, Professor Aranowska did not live to see this book published, but her spirit and contribution are present on every page. I will refer many times to our shared ideas, collaborative works, and conference presentations, which were key to my scientific development.

I would also like to sincerely thank Dr Jolanta Rytel for many years of discussions on statistical, methodological, and psychometric topics. Dr Rytel was a guide to me, particularly at the beginning of my career, but also later on, when her support and expertise were extremely valuable. I am grateful for all her help and for the fact that I could always rely on her substantive assistance in solving various problems.

My special thanks also go to Professor Czesław Nosal for undertaking the review of this book, for his exceptionally insightful comments, and for the immense support and kindness he offered during the preparation of the manuscript. His academic thoroughness and friendly approach were for me not only a great help, but also a source of deep encouragement in finalising this publication.

I would also like to express special thanks to Professor Lidia Grzesiuk, Professor Barbara Bokus, Professor Ewa Rzechowska, and Professor Waldemar Koczkodaj

for their collaboration and contributions to the development of artificial intelligence applications and other algorithms in psychometrics and psychology. As mentioned in the preface, each of these individuals had a significant influence on my development and on the implementation of modern methods in psychology. I thank them for their support and for the opportunity to work together on such important topics.

I am also deeply grateful to my friends and colleagues: Professor Tomasz Prusiński, Dr Joanna Świderska, Professor Monika Młynarska, and Dr Kamila Dobrenko. Thank you for the shared conference trips, the time spent together, the scientific discussions, and for every form of support I received in various aspects of my work. Your presence and encouragement were invaluable to me.

Finally, I extend my heartfelt thanks to my students–doctoral candidates from the Institute of Psychology at UKSW, students of the Big Data in Social Analytics programme at Cardinal Stefan Wyszyński University, Erasmus students visiting UKSW, and all students I had the pleasure of teaching at other institutions. Thank you for the discussions we had, which often inspired further exploration of new problems and significantly contributed to the creation of this book.

Agnieszka Szymańska

Warsaw, May 2025

INTRODUCTION

Mathematical modeling is the process of creating abstract, mathematical representations of real-world phenomena or processes for the purpose of analysis, prediction, and understanding. In the natural and technical sciences, mathematical modeling serves as a fundamental tool that enables researchers to transform complex, often multidimensional problems into more comprehensible and analytically tractable mathematical forms (Lee & Buzby, 2021; Rymanowski, 2007).

A mathematical model consists of mathematical equations that describe phenomena in a quantitative manner, allowing for formal analysis and prediction of their behaviour. In mathematical modeling, the application of appropriate mathematical methods—such as numerical methods, linear algebra, functional analysis, or probability theory—is essential to solving these equations (Lee & Buzby, 2021).

The process of mathematical modeling begins with the identification and understanding of the problem, along with the specification of key variables that exert the greatest influence on the phenomenon to be modelled. Based on these variables, appropriate mathematical equations are formulated to form the foundation of the model. This model is then tested and verified using available empirical data, allowing it to be refined so that it represents reality as accurately as possible.

Assumptions also play a significant role in mathematical modeling and constitute an inherent component of every model. These assumptions refer to simplifications of reality that are necessary for constructing the model but may also affect its accuracy and predictive power. For this reason, the process of mathematical modeling does not end with model construction but also includes testing and interpretation of results within the context of the model's assumptions and limitations (Lee & Buzby, 2021).

In the context of psychological sciences, mathematical modeling is becoming increasingly widespread, enabling researchers to understand complex psychological phenomena—such as cognitive, emotional, social, and behavioural processes—in a more quantitative and formal manner. In particular, mathematical modeling makes it possible not only to describe these processes but also to predict their course and analyse the influence of various factors on individual behaviour.

However, mathematical modeling in psychology requires a specific approach that takes into account both the complexity of the human psyche and the limitations associated with psychological measurement. Therefore, in order for mathematical models to be useful in psychology, they must be not only precise but also flexible and capable of adapting to diverse research contexts (Aranowska, 1996, 2005).

Mathematical modeling is a fundamental tool used across scientific disciplines to understand complex phenomena and processes. In psychology—which investigates human behaviour, emotions, and cognitive processes—mathematical modeling plays a particularly important role, enabling researchers to apply a systematic and quantitative approach to the analysis of psychological data.

However, the specific nature of psychology presents particular challenges for mathematical modeling. Most notably, psychology operates with a wide range of data types—from nominal scales and ordinal data to interval and ratio scales. This diversity necessitates the use of specialised analytical methods that allow for the appropriate processing and interpretation of results (Aranowska, 2005).

It is also important to highlight the distinctive character of psychological data. Unlike more stable scientific fields such as physics or engineering, psychology measures both enduring traits (e.g. personality traits) and mental states, which tend to fluctuate considerably. This poses a challenge, as human beings are subject to internal and external influences that can alter measurement results within a short period of time. Consequently, psychologists must contend with greater measurement error stemming from the nature of the human psyche. Achieving stable measurement results is more difficult, which requires mathematical modeling to place particular emphasis on validation and the assessment of model reliability and construct validity (Aranowska, 2005).

Moreover, many of the constructs studied in psychology are latent—meaning they are not directly observable. Latency implies that we must rely on indirect indicators to infer the presence of characteristics such as intelligence, anxiety, or empathy. These hidden traits pose an additional challenge for mathematical modeling, which must take into account both intra-individual and inter-individual variability (Aranowska, 2005).

In psychology, mathematical modeling serves not only as a tool for data analysis but also as a means of gaining theoretical understanding of psychological phenomena. Traditional models, such as structural equation models (SEM), allow researchers to test psychological theories and uncover new relationships between variables (Konarski, 2009; Szymańska, 2016b). However, as psychology continues to develop,

there is a growing need for more advanced modeling methods capable of capturing the complexity and multidimensionality of the phenomena under investigation.

In this context, the role of artificial intelligence (AI) is becoming increasingly significant. AI enables the analysis of vast data sets, the discovery of hidden patterns, and the modeling of psychological phenomena that would be difficult to capture using traditional methods. AI can support psychologists not only in data analysis but also in the construction of new tests, the modeling of latent traits, and the validation of theoretical models.

This book discusses a variety of methods and techniques for mathematical modeling in psychology, with particular emphasis on the use of artificial intelligence algorithms. Both traditional approaches and innovative solutions will be presented—solutions that have the potential to revolutionise the way psychology analyses data and develops theories. We will also address the challenges arising from the nature of psychology as a science, including difficulties related to measurement error and the modeling of latent traits.

Psychometrics, as a field concerned with the theory and techniques of measurement in psychology, plays a key role in the development and validation of psychological tests. Traditional approaches to psychometrics are based on classical test theory (CTT) and item response theory (IRT), which provide tools for assessing the validity and reliability of psychological tests (Anastasi & Urbina, 1999; Hornowska, 2003). However, the development of artificial intelligence (AI) algorithms opens new possibilities in this field, allowing for more advanced approaches to psychometric analysis.

Artificial intelligence can support psychometric processes at multiple levels. First, AI can be used for the automatic generation and calibration of psychological tests. Using algorithms such as decision trees, neural networks, or genetic algorithms, it is possible to create dynamic tests that adapt to respondents' answers, maximising the validity and reliability of the measurement (Szymańska, 2024b). Second, AI can assist in identifying and eliminating errors in psychological tests—such as item bias or sequence effects—thus improving the overall quality of psychometric tools.

One of the most promising applications of AI in psychometrics is the analysis of psychological test results. Traditionally, the analysis of results has relied on simple statistical models, which may not have captured the full complexity of psychological data. Thanks to its capacity to process large data sets and uncover hidden patterns, AI enables a more comprehensive analysis of test results, leading to more accurate diagnoses and a deeper understanding of the studied phenomena (Szymańska, 2024b).

AI can also support psychometricians in the process of test validation by verifying whether a measurement tool truly measures what it is intended to, and whether it can accurately differentiate between various groups of respondents. Employing AI in the validation of psychological tests, in terms of both reliability and validity, can accelerate the development and refinement of diagnostic instruments,

which is particularly important in fields where fast and precise diagnosis is critical (Szymańska, 2024b).

Before we move on to the discussion of AI applications in psychology, it is worth reflecting on how AI has been used in this field in the past. Early attempts to apply AI in psychology date back to the 1960s, when the first expert systems were developed—such as ELIZA, which simulated basic processes of psychotherapy (Wikipedia contributors, 2024). Although ELIZA was a relatively simple program, its creation inspired researchers to further experiment with AI in psychological contexts.

In the following decades, the advancement of computers and AI algorithms led to more sophisticated applications in psychology. Expert systems began to be used in diagnostics, psychological data analytics, and the modeling of cognitive processes. One example is the MYCIN project, which not only diagnosed diseases based on symptoms but also provided treatment recommendations—a similar approach could be adopted in psychology to develop systems supporting psychological diagnoses (Wikipedia contributors, 2024).

Thanks to the application of AI, researchers were able for the first time to systematically analyse vast data sets that had previously been beyond their reach—contributing to the development of the Big Data era (Ptaszek, 2019; Stephenson, 2018). AI has also facilitated the emergence of new research methods, such as text analysis, emotion recognition in speech and facial expressions, and the modeling of social interactions. This, in turn, has opened up new research opportunities, allowing for a more precise and comprehensive understanding of human behaviour.

Mathematical modeling in psychology, supported by AI algorithms, opens up new horizons, enabling more precise and complex analyses that may lead to a deeper understanding of the human psyche. As tools and techniques of artificial intelligence continue to evolve, psychology as a science gains new research capabilities that have the potential to revolutionise how we perceive and analyse psychological phenomena. The future of psychology will undoubtedly be closely linked to the continued development of artificial intelligence, and this book aims to present what is already possible today—as well as what may become reality in the near future.

PART I

Latent Variable Models and Structural Equation Modeling

CHAPTER 1

Characteristics of Models Verified by Means of Structural Equation Modeling

1.1. Introduction to Structural Equation Models

Structural Equation Modeling (SEM) is an advanced statistical method used for the empirical verification of scientific theories. This method is particularly important in the context of contemporary research methodology in the social sciences and psychology, where the accuracy of representing theoretical phenomena in models is crucial for their understanding and the further development of theory (Aranowska, 2005; Hair et al., 2006; Konarski, 2009; Szymańska, 2016b).

A theory, as an ordered system of scientific laws, explains a specific type of phenomena according to a unified principle. Its purpose is not only to describe but also to explain structure and processes, which enables the prediction of future phenomena. A key aspect of theory is its ability to determine the relations between constituent elements, which requires a detailed analysis of initial and structural conditions (Nowak, 2007). Therefore, the empirical verification of theoretical models constitutes an indispensable stage that allows for the assessment of the extent to which the model's structure reflects the assumptions of the theory and to what degree empirical data confirm the validity of this representation.

The structural reconstruction of a theory is a fundamental process in SEM. It involves the analysis of theoretical material, including the scopes of fundamental concepts, laws, and principles that constrain the functioning of the theory (Jonkisz, 1998). An essential step in this process is identifying the structure of the theory and its connections with other theories, and then presenting it in the form of a model.

This model, often represented by means of mathematical graphs, visualizes the theoretical structure and enables the verification of the theory's assumptions.

A key role in the reconstruction of theory is played by the concept of the model as an interpretation of the axioms of the theory. The model, as a structural reconstruction of the theory, represents theoretical assumptions in a systematic and logical manner, which enables the empirical verification of the relations between the elements of the theory. It is worth emphasizing that various types of relations, such as association, mediation, or moderation, play a key role in structural analysis (Suppes, 1972).

During the reconstruction of a theory, it is essential to take into account the scope and characteristics of its structure, which is clearly presented in Figure 1.1. The structure of a theory can be described as a system in the form <Xj, Rj>, where Xj is a sequence of elements that constitute the structure, also referred to as *the scope of the structure*, and Rj represents the relations between these elements, that is, the characteristics of the *structure's scope* (Szymańska, 2016b). In SEM, we distinguish several types of relations between structural elements, which serve various analytical roles:

The relation of association (denoted as R3.1 in Figure 1.1) refers to a connection between variables without specifying a clear causal direction, suggesting co-occurrence or correlation. In diagrams, it is represented by a curved arrow with arrowheads on both ends.

Deterministic relations (denoted as R2.1, R4.2, R4.3 in Figure 1.1) refer to direct connections between variables, symbolized by straight arrows with unidirectional heads.

Mediation relations (denoted as R2.1, R4.2 in Figure 1.1) occur when one variable mediates the relationship between two other variables, enabling the explanation of the mechanism of their mutual influence.

Moderation relations (denoted as R5.4, R6.5, R4.6 in Figure 1.1) concern situations in which the strength or direction of the effect of one variable on another depends on a third variable serving as a moderator. These relations may lead to synergy or buffering effects and are represented in SEM graphs by straight arrows with a unidirectional head.

Synergy and buffering relations refer to the manner in which moderating variables alter the relationship between two other variables.

The synergy effect occurs when the presence of a third variable (moderator) strengthens or amplifies the impact of one variable on another (Nowak, 2007). For example, if variable A has a positive effect on variable B, and the inclusion of moderator C increases this effect, we speak of a synergy effect. In such a case, the moderating variable causes the effect to become stronger than it would be in its absence.

The buffering effect takes place when the presence of the moderator weakens or reduces the impact of one variable on another (Nowak, 2007). An example might be a situation where variable A negatively affects variable B, but the introduction of moderator C reduces this negative impact. The buffering relation thus functions protectively, alleviating or diminishing the effect of variable A on variable B.

In both cases, the third variable plays a key role in modifying the strength or direction of the relationship between the two other variables, allowing for a better understanding of the mechanism of their mutual interactions.

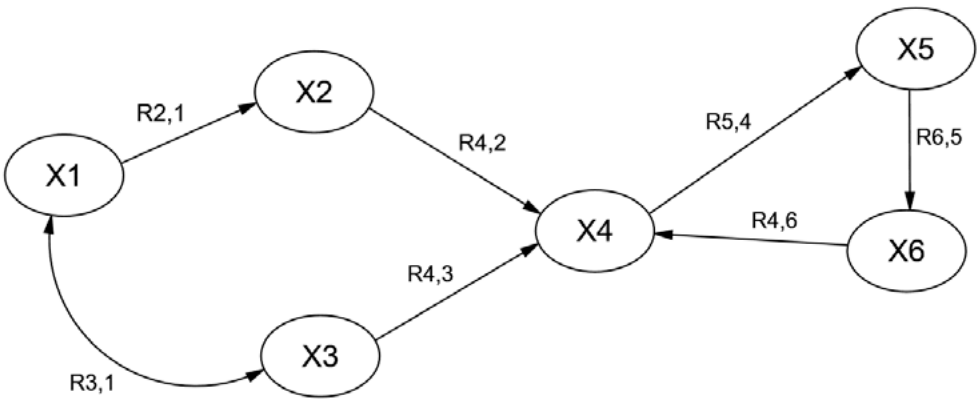


Figure 1.1. Scope and Characteristics of the Structure

The figure presents a schematic arrangement of relations between variables in an SEM model. The meaning of individual variables and the interpretation of the relations between them depend on the content assigned to them during the process of operationalisation.

In the process of theory reconstruction, it is also necessary to take into account all key variables in order to avoid simplifications that may lead to an insufficient description of the phenomenon. In this context, it is particularly important to identify exogenous variables, which initiate processes within the model, as well as endogenous variables, which are dependent on them (Hair et al., 2006).

The precise specification of these variables allows for the construction of a model that not only verifies the theory but also reconstructs the processes under analysis. Exogenous variables serve the role of independent variables, while endogenous variables are dependent both on the exogenous variables and on one another. It is also worth drawing attention to the issue of the role of theory in describing phenomena. In SEM, theory is not merely a set of hypotheses but a structural model that specifies precise relations between variables. It is important to emphasize that these relations are empirical in nature and reflect dependencies observed in the data.

In SEM, theory not only explains phenomena but also predicts their course in new contexts. However, as a statistical method, SEM cannot directly prove the existence of causal relationships, which traditionally require experimental studies (Garson, 2013). SEM does, however, allow for the probabilistic support of such relationships, enabling the testing of whether empirical data are consistent with predictions derived from a theoretical causal model. In this way, SEM supports research on causal relationships by providing tools for analysing complex phenomena, even if it is not capable of delivering direct evidence of causality (Szymańska, 2016b; Szymańska & Aranowska, 2016).

The reconstruction of a theory therefore requires the inclusion of both a broad range of concepts and precise relations among them. This process, in line with the philosophical concept of theory axiomatization, gives the theory a structure that is not only coherent but also empirically verifiable (Suppes, 1972).

Ultimately, the structural model becomes a substantial representation of the theory, whose purpose is not only to explain but also to predict and forecast phenomena based on clear and logical relations between its elements. However, the effectiveness of such a representation depends on the accuracy of the theoretical foundations and the correctness of variable operationalisation, as the incorrect construction of indicators may lead to a distorted image of the structure, despite a good model fit to the data.

1.2. Overview of Key Concepts and Definitions

The operationalization of theoretical concepts constitutes a fundamental stage in the process of verifying SEM (Structural Equation Modeling) models. This process requires a precise definition of the conceptual scope, that is, the connotation of the terms, which makes it possible to identify the set of all referents of a given term, thereby enabling the specification of its denotation. Operationalization involves the need to develop measurement tools that reliably and validly reflect the essence of the studied phenomenon, allowing for its accurate analysis and interpretation (Nowak, 2007).

In the context of psychological and social sciences, the studied properties often have a latent nature, meaning they are not directly observable and are inferred based on observable indicators. An indicator is a measurable manifestation of the studied construct that indirectly informs us of its presence. The studied construct, also referred to as the *indicatum*, relates to the theoretical attribute that one aims to measure by means of indicators. In this context, a key task is to select indicators appropriately, as they determine the reliability and validity of the measurement (Nowak, 2007).

The fundamental features of a good indicator are its inclusion power and exclusion power. *Inclusion power* refers to the indicator's ability to encompass everything that pertains to the studied construct (indicatum), while *exclusion power* refers to the indicator's ability to exclude everything that does not belong to that construct (Nowak, 2007). An ideal indicator, characterized by maximum inclusion and exclusion power, would represent a pure marker, meaning it would contain everything related to the studied construct and simultaneously exclude everything unrelated to it. Moreover, such an indicator would exhibit maximum correlational power, i.e., it would always appear when the indicatum is present, thus ensuring full coverage of the studied trait in measurement.

Unfortunately, in practice, achieving a perfect indicator is an extremely difficult task. Therefore, in most cases, multiple indicators are employed in order to best capture the complexity of the studied construct (Nowak, 2007). In the context

of SEM, the number of indicators is of key importance, as it affects the complexity of the model, which in turn determines the research procedures that must be applied. For this reason, proper operationalization of concepts, selection of appropriate indicators, and precise specification of exogenous and endogenous variables constitute the foundation of effective SEM analysis.

Diligence at these stages is crucial, as any errors made during operationalization may have far-reaching consequences for the entire model. Improper operationalization of concepts leads to model uncertainty, and its results may be misleading or simply incorrect (Nowak, 2007). If the indicators do not fully reflect the meaning of the studied construct (indicatum) or contain elements that do not belong to it, this results in reduced reliability and validity of the model. As a consequence, the SEM model may fail to fulfil its fundamental task, which is theory verification. This can lead to the rejection of a correct theoretical model due to poor fit with improperly selected indicators, or to the acceptance of an incorrect model that merely appears to fit the empirical data.

Faulty operationalization of concepts also affects the model's ability to predict and verify. A model that lacks reliability and validity cannot provide credible forecasts, and its assumptions cannot be effectively tested. This means that even if the theoretical foundation of what is to be examined is sound, failure to operationalize it correctly may result in unsuccessful model verification at the stage of SEM analysis (Szymańska, 2016b). The consequence of faulty operationalization is the risk of losing the scientific value of the entire study. The results become ambiguous or misleading, which may compromise the entire research process. Therefore, it is crucial that indicators not only accurately represent the studied construct, but also demonstrate validity and reliability, allowing for sound inferences about the phenomena under investigation. Without these foundations, the SEM model loses its explanatory and predictive power, which in extreme cases may lead to false conclusions and erroneous research decisions (Szymańska, 2016b).

The concepts of exogenous and endogenous variables also play a key role in SEM. *Exogenous variables* are independent variables that initiate processes within the model, causing changes in the values of endogenous variables through specified directional relationships. *Endogenous variables* are thus dependent on exogenous variables as well as on mutual relationships between themselves. The precise identification of these variables is crucial for constructing a model that not only verifies a theory but also enables the reconstruction of the processes under analysis.

In summary, a well-considered and accurate operationalization of concepts, appropriate selection of indicators, and precise definition of exogenous and endogenous variables constitute the foundation for the effective application of SEM in scientific research. By maintaining diligence at these stages, it becomes possible to obtain reliable results that contribute to the development of both theory and practice in the social and psychological sciences (Szymańska, 2016b).

CHAPTER 2

First-Order Measurement Model

2.1. Definition and Applications

Structural Equation Modeling (SEM) is a technique composed essentially of two stages: the construction of the measurement model and the estimation of the full structural model (Bartholomew et al., 2008; Hair et al., 2006; Szymańska, 2016b). This chapter focuses on the construction of the first-order measurement model, which represents the first step in theory verification using SEM. It discusses the issue of selecting items for latent structures and the principles of verifying the measurement model through Confirmatory Factor Analysis (CFA).

The first-order measurement model is a key component in SEM analyses, as it defines the relationships between observable variables and latent (hidden) variables. In psychology, where the studied characteristics are typically latent—that is, not directly observable—the importance of the measurement model is particularly pronounced (Aranowska, 2005). A psychologist does not directly observe the object of study, such as a person's mental traits, but instead infers their presence based on observable indicators, such as questionnaire responses. In this context, observable variables serve as indicators that enable the inference of latent traits—i.e., the theoretical construct intended to be measured (Aranowska, 2005).

The construction of the measurement model is the first step in the verification of models using SEM. A crucial task is to ensure that the operationalization of latent variables has been correctly conducted. Observable variables must be valid and reliable in relation to the latent variables they represent. Validity refers to the extent to which the observable variables truly reflect the constructs they

are intended to measure, while reliability pertains to the stability and consistency of those measurements across various conditions (Aranowska, 2005).

Path analysis, as an earlier approach to modeling relationships between variables, focused on observable variables. However, in psychology—where the objects of study are inferential (latent) traits—it is necessary to model variables that are not directly observable. This is what distinguishes the SEM approach from earlier techniques: the ability to incorporate latent variables into the model, allowing for a more accurate representation of complex psychological relationships (Hair et al., 2006).

The first-order measurement model in SEM enables the assessment of how accurately observable variables represent hidden traits. The proper construction of the measurement model is a necessary condition for obtaining reliable SEM analysis results. If the operationalization is conducted carelessly, the entire model loses stability, which may lead to incorrect conclusions about relationships among variables in the structural model (Szymańska, 2016b).

In practice, the first stage of SEM modeling involves building and testing the measurement model, which is equivalent to Confirmatory Factor Analysis (CFA). This means that before estimating the relationships between latent variables in the full structural model, one must verify whether the indicators truly represent the assumed theoretical constructs. Only after confirming this relationship through CFA can one proceed to estimate the paths between constructs within SEM.

CFA serves to verify whether the selected indicators indeed measure the assumed latent constructs, which is a necessary step before conducting further structural analyses. Verification of the measurement model enables the identification of potential issues related to the operationalization of variables, which is crucial for the validity of the structural model analysis (Bartholomew et al., 2008; Hair et al., 2006; Szymańska, 2016b).

In summary, the first-order measurement model is the foundation upon which further SEM analysis is built. Its proper construction and verification determine the validity and reliability of the entire model's results, which directly impacts the quality of the scientific conclusions drawn from the study. Consequently, SEM analysis can provide more precise insights into causal relationships and the theoretical structure of the studied phenomena, contributing to the development of theory and practice in the social and psychological sciences.

2.2. Formal Assumptions of Measurement Models

The following presents a general procedure for constructing and validating structural equation models. The generality of the description stems from the considerable mathematical complexity of the method, the details of which go beyond the scope of this work (Hair et al., 2006).

In the first stage of model validation, a measurement model is constructed and subsequently verified using *Confirmatory Factor Analysis* (CFA). The purpose of

building the measurement model is to assess whether the latent constructs have been properly specified—that is, whether a latent factor sufficiently explains the variance of its observable variables[1].

The structure of an observable variable's outcome consists, in accordance with the linear representation in Equation 2.1, of several components:

$$\chi_1 = \lambda_{x1,i} + \sigma_1, \tag{2.1}$$

where: λ denotes the factor loading, σ represents the error variance, χ symbolizes the observable variables, and ξ the latent constructs. The measurement model, like the structural model, can be represented using graphs. An example of a measurement model is shown in Figure 2.1. A characteristic feature of graphs describing the measurement model is the bidirectional relationship between latent constructs (a double-headed arrow), which symbolizes the correlation between these latent variables, denoted by the Greek letter ϕ (phi).

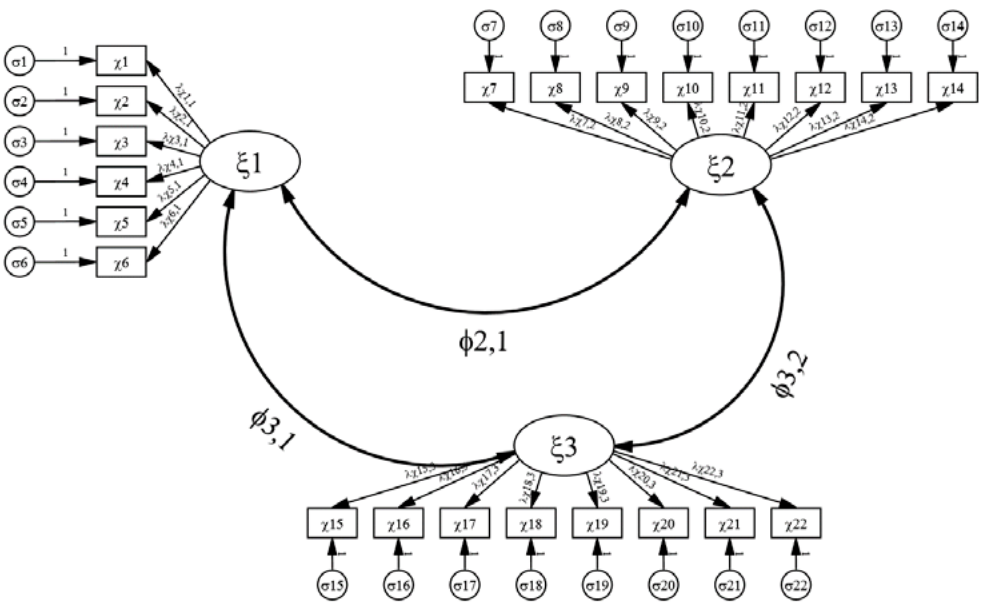


Figure 2.1. presents a graph illustrating the measurement model, where λ denotes the factor loading, σ represents the error variance, χ symbolizes the observable variables, ξ denotes the latent constructs, and ϕ represents the correlations between constructs

The figure depicts the schematic arrangement of dependencies within the measurement model in SEM. The interpretation of individual graphical elements, as well as the meaning of relationships between variables, depends on the operationalization adopted for the latent construct.

[1] It should be noted that it is the latent structure that explains the variability of the observable variables, not the other way around.

Since correlations exist between latent variables, latent variables are not assigned measurement errors. Furthermore, it is assumed that all latent variables are interconnected, meaning that each should be correlated with every other latent variable (Hair et al., 2006; Szymańska, 2016b).

In the case of observable variables, it is assumed that their measurement errors should not be correlated unless there are strong theoretical or empirical justifications for doing so. In the graphs of a measurement model, observable variables are represented as rectangles marked with χ, while their measurement errors are depicted as small circles labeled with the Greek letter sigma (σ). Measurement errors refer to the discrepancies between the true value of a variable and its measured value for observable variables. Correlating measurement errors without appropriate justification violates the model's assumptions and may lead to distorted analysis results and incorrect interpretations (Hair et al., 2006; Szymańska, 2016b).

A strong assumption in modeling is that each observable variable should be assigned exclusively to a single latent variable (Hair et al., 2006).

When constructing a measurement model, particular attention must be paid to five key principles that apply to every model of this kind (Hair et al., 2006).

The first principle concerns the values of factor loadings (λ), which should exceed 0.7. This is the minimum acceptable value because its square equals 0.49, meaning that the latent construct explains at least 50% of the variance in the observable variable, while the remaining 50% constitutes error (σ)—the unexplained variance. When the proportion of error begins to dominate, the value of the observable variable as an indicator of the construct becomes questionable. Therefore, a value of 0.7 is considered critical for the lambda coefficient.

The second principle refers to the necessity of monitoring the *variance extracted* (VE) for each latent variable. The VE value serves as a measure of how well the observable variables explain the variance of their latent variable. It is calculated according to Equation 2.2:

$$\text{VE} = \frac{\sum_{i=1}^{n} {\lambda_i}^2}{n} = \frac{\sum_{i=1}^{n} (\lambda_i - 0)^2}{n}, \tag{2.2}$$

where i denotes the index of the observable variable serving as an indicator of the latent variable.

This principle emphasizes that the values of factor loadings (λ_i) must be sufficiently high for VE to exceed 0.5. Otherwise, if VE is lower, it indicates that a substantial portion of the variance in the observable variables remains unexplained by the latent variable, rendering the observable variables unreliable indicators of the construct. This undermines the validity of the measurement model.

This principle is directly linked to the first one, as high factor loadings are a necessary condition for achieving an adequate level of VE and for recognizing that the construct properly represents its indicators.

The third principle concerns the need to calculate the reliability of each latent construct in the measurement model. This reliability, denoted as CR (construct

reliability), is a measure of the internal consistency of the indicators assigned to a given construct. It is calculated according to Equation 2.3 (Hair et al., 2006):

$$\text{CR} = \frac{\left[\sum_{i=1}^{n} \lambda_i\right]^2}{\left[\sum_{i=1}^{n} \lambda_i\right]^2 + \left[\sum_{i=1}^{n} \sigma_i\right]}, \tag{2.3}$$

where CR denotes the construct reliability, λ_i are the factor loadings of the observable variables, and σ_i are the error variances associated with these variables.

According to this principle, the CR value should be at least 0.6, with values in the range 0.6 < CR < 0.7 considered acceptable, and values greater than 0.7 indicating good construct reliability.

However, it is worth noting that this formula, despite its popularity, has certain significant limitations. One such limitation is that its value may increase with the number of observable variables (e.g., test items), meaning that adding more indicators to the model can artificially inflate the reliability estimate. In other words, the more items— even randomly selected—are added to the construct, the higher the CR value becomes. This mechanism is similar to that of Cronbach's alpha, which is also sensitive to the number of items and can lead to an overestimation of reliability.

Additionally, it can be argued that the formula sums the correlations before squaring them. The coefficient may thus suggest greater internal consistency of the construct than is actually justified, potentially leading to incorrect conclusions about the validity of the measurement model.

The numerical example below demonstrates that the classical CR formula may overestimate the reliability of a construct, as it adds the factor loadings before squaring them, thereby mathematically incorporating the correlations between indicators. Let us assume that we have three observable variables with the following factor loadings:

$$\boldsymbol{\lambda_1 = 0.3, \lambda_2 = 0.5, \lambda_3 = 0.6}$$

Construct reliability (CR) is a measure of how well the observable variables reflect the latent variable. We proceed in accordance with the standard method for calculating the CR formula.

Step 1: Calculate the sum of factor loadings

$$\boldsymbol{\Sigma\lambda_i = 0.3 + 0.5 + 0.6 = 1.4}$$

Step 2: Square the sum of factor loadings

$$\boldsymbol{(\Sigma\lambda_i)^2 = 1.4^2 = 1.96}$$

Step 3: Calculate the error variance for each observable variable

$$\sigma_1 = 1 - 0.3^2 = 0.91$$
$$\sigma_2 = 1 - 0.5^2 = 0.75$$
$$\sigma_3 = 1 - 0.6^2 = 0.64$$

Step 4: Calculate the sum of error variances

$$\Sigma\sigma_i = 0.91 + 0.75 + 0.64 = 2.3$$

Step 5: Calculate construct reliability (CR)

$$\mathbf{CR} = \frac{1.96}{1.96 + 2.3} = \frac{1.96}{4.26} = 0.46.$$

The resulting construct reliability is approximately 0.46, which is below the acceptable threshold of 0.7. Such a result indicates that the observable variables poorly reflect the latent variable, suggesting that the variable may not be sufficiently reliable for research purposes. The low CR value may result from low factor loadings or high measurement errors, which points to the need to reconsider the selection of observable variables or the structure of the measurement model.

Next, we apply a correction by first squaring each factor loading and then summing them.

Given lambda values:

$$\lambda_1 = 0.3, \lambda_2 = 0.5, \lambda_3 = 0.6.$$

Revised calculation method:

Step 1: Square the factor loadings

$$\lambda_1^2 = 0.32 = 0.09$$
$$\lambda_2^2 = 0.52 = 0.25$$
$$\lambda_3^2 = 0.62 = 0.36$$

Step 2: Sum the squared loadings

$$\textbf{Sum of squared lambdas} = 0.09 + 0.25 + 0.36 = 0.70$$

Step 3: Calculate the error variances for each observable variable and sum them

$$\sigma_1 = 1 - 0.3^2 = 0.91$$

$$\boldsymbol{\sigma_2 = 1 - 0.5^2 = 0.75}$$

$$\boldsymbol{\sigma_3 = 1 - 0.6^2 = 0.64}$$

$$\boldsymbol{\Sigma\sigma_i = 0.91 + 0.75 + 0.64 = 2.3}$$

Step 4: Calculate construct reliability (CR)

$$\mathbf{CR = \frac{0.70}{0.70 + 2.3} = \frac{0.70}{3.00} = 0.2333.}$$

This result shows that using the revised method (squaring the loadings first and then summing them) produces a much lower CR value than the original calculation. This highlights the difference between summing the squares and squaring the sum.

The author therefore proposes using a more precise formula for calculating the reliability of a latent variable developed by Aranowska, which constitutes a correction to the traditional CR formula (Szymańska & Aranowska, 2016).

(2.4)
$$\gamma = \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{\sum_{i=1}^{k} {\lambda_i}^2 + \sum_{i=1}^{k} \left(1 - {\lambda_i}^2\right)}} = \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{\sum_{i=1}^{k} {\lambda_i}^2 + \sum_{i=1}^{k} 1 - \sum_{i=1}^{k} {\lambda_i}^2}} = \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{k}}$$

where:
γ is an abbreviation for *construct reliability*,
i is the index of the observable variable serving as an indicator of the latent variable,
λ denotes the factor loading,
δ represents the error variance.

The Aranowska γ-coefficient enables a more accurate estimation of construct reliability by correcting for the methodological limitations of the classical CR formula. It is also resistant to the number of items—its value does not artificially increase when more weak indicators are added, which is a common issue for some traditional reliability indices. Moreover, with a small number of items showing high lambda values (factor loadings), this coefficient still retains high reliability, whereas the CR coefficient tends to show lower reliability in such cases. As the number of items increases, the CR coefficient also increases its reliability—even when factor loadings are weak—while the Aranowska coefficient remains unaffected by the number of items, avoiding artificial inflation of its value.

It should also be emphasized that introducing this correction was necessary, as some reliability formulas found in the literature mistakenly suggest summing correlations before squaring them, which is not methodologically justified—correlations should not be added.

The fourth principle concerns the need to ensure discriminant validity. In a measurement model, it is required that observable variables are explained exclusively by their own latent construct and are not simultaneously explained by other latent variables. This means that cross-loading—i.e., partial content overlap between different constructs—should be avoided. Discriminant validity is crucial, as it allows one construct to be clearly distinguished from another.

An alternative approach to assessing discriminant validity is to compare the factor loadings (λ) with the correlation between latent variables (φ). If the factor loadings are higher than the φ correlation, it indicates that the given construct explains the variance of its observable variables better than any other construct. This confirms the greater discriminant power of those observable variables in relation to other constructs (Hair et al., 2006).

This principle highlights that correctly assigning observable variables to the appropriate latent constructs is essential for the proper functioning of the model and for avoiding erroneous conclusions in SEM analysis.

The fifth principle concerns the evaluation of the measurement model's fit using fit indices such as χ^2 and RMSEA. When constructing a measurement model, it is essential to ensure that the values of these statistics indicate a good fit between the model and the empirical data.

The χ^2 test is one of the most important tools for assessing model fit. Ideally, the χ^2 statistic should be low enough not to exceed the critical value in the χ^2 probability distribution corresponding to a significance level of 0.05 with a given number of degrees of freedom (as defined in Equation 2.4). This means that the p-value should be greater than 0.05, indicating no significant differences between the measurement model and the observed data. However, it is worth noting that the χ^2 test is sensitive to sample size and the number of degrees of freedom. With large samples, even minor deviations from the model can result in a significant χ^2 value, despite the model being well-fitted in practice. For this reason, it is advisable to supplement the model fit assessment with additional indicators.

One such indicator is RMSEA, which evaluates model fit while accounting for the number of degrees of freedom. An RMSEA value below 0.08 is generally considered to indicate good model fit. However, according to the latest standards, an RMSEA value below 0.06 is more desirable. The lower the RMSEA value, the better the model fit. It is also important to emphasize that RMSEA is sensitive to sample size, and therefore achieving low RMSEA values in studies with small samples may be more challenging.

It should also be noted that model fit evaluation should not rely solely on a single statistic (Konarski, 2009). It is recommended to use multiple fit indices, such as the CFI (Comparative Fit Index) and TLI (Tucker–Lewis Index), which, together with χ^2 and RMSEA, provide a more comprehensive picture of model fit. For example, a CFI value greater than 0.90 and a TLI value greater than 0.90 are considered to indicate good model fit (Bartholomew et al., 2008; Hair et al., 2006).

Moreover, the standards for evaluating model fit evolve alongside scientific progress. For instance, the RMSEA value once considered acceptable has changed

over time—from below 0.10, to 0.08, and now to the currently recommended threshold of below 0.06. Therefore, it is essential to regularly monitor updated guidelines and standards in the scientific literature to ensure that the fit criteria being applied are consistent with the most recent recommendations.

In conclusion, adhering to this principle during the construction of a measurement model, as well as regularly updating one's knowledge of fit statistics standards, is crucial for ensuring that the model is not only well-fitted to the data but also scientifically justified and aligned with current methodological guidelines.

Degrees of freedom for the χ^2 statistic in a CFA model are calculated according to Equation 2.5:

$$df = q(q + 1)/ 2 - p, \tag{2.5}$$

where q denotes the number of all observable variables, and p represents the number of estimated parameters in the model.

Once the measurement model has been successfully constructed, the next step is the validation of the full structural equation model. This issue will be discussed in the following chapter.

2.3. Example of a First-Order Measurement Model

As previously mentioned, prior to conducting the study it is necessary to construct a theoretical model that specifies the detailed relationships between latent constructs.

The measurement model was developed to verify whether the latent constructs had been properly specified. Based on theoretical assumptions, three latent variables were constructed:

(a) **Discrepancy** (see "discrepancy" in Figure 2.2), a variable reflecting the gap between the parental goal (i.e., the trait the parent wishes to develop in the child) and the child's current level of development in that area. This refers to a cognitive assessment of the mismatch between expectations and the educational reality;

(b) **Experienced parental difficulty** (see "difficulty" in Figure 2.2), defined as the parent's subjectively perceived level of tension, overload, or frustration associated with raising the child. This variable does not refer directly to the child's behavior but rather to the psychological cost incurred by the parent in their everyday parenting functioning;

(c) **Representation of the child** (see "representation" in Figure 2.2), understood as the internal image of the child in the parent's mind, shaped by both past and present interactions. It includes both cognitive and emotional components, influencing how the parent interprets the child's behavior and makes parenting decisions accordingly.

The graph presenting the results of the Confirmatory Factor Analysis is shown in Figure 2.2.

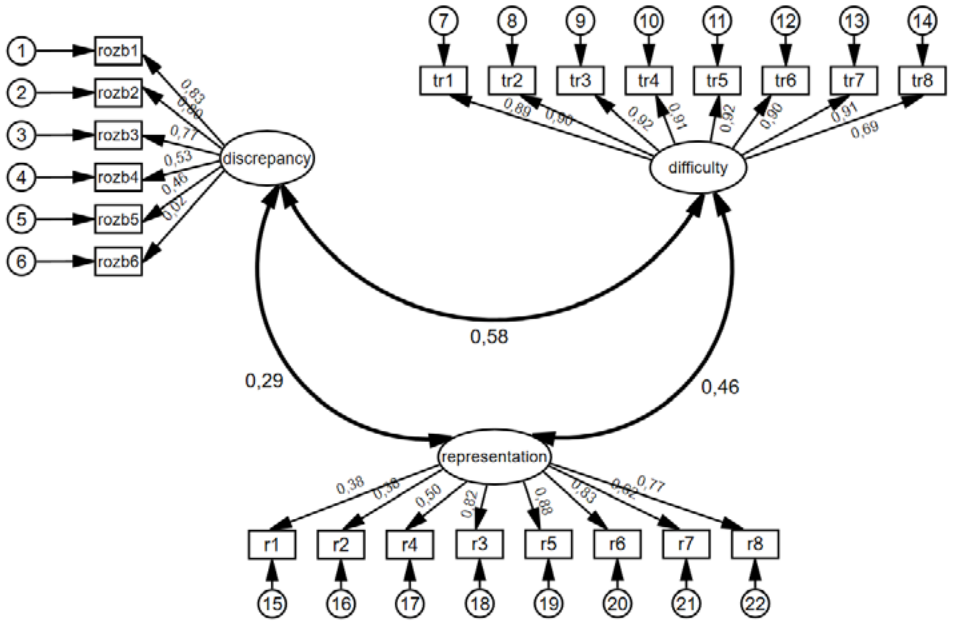


Figure 2.2. Graph presenting the results of the Confirmatory Factor Analysis model

The latent structures of the variables were defined as unidimensional, despite the multifactorial nature revealed in the Exploratory Factor Analysis. This is a single-level Confirmatory Factor Analysis model, which was deliberately applied to illustrate the consequences of disregarding the complexity of latent structures.

Below is a characterization of each variable.

Discrepancy was operationalized using the Discrepancy Scale, which served to assess the degree of mismatch between the intended parental goal and the current level of the child's development in a given trait. This construct is cognitive in nature and refers to the perception of the difference between the parent's educational plan and their observation of reality.

The factor loadings (lambdas) of the observed variables are as follows: $\lambda_{rozb1} = 0.830$; $\lambda_{rozb2} = 0.800$; $\lambda_{rozb3} = 0.770$; $\lambda_{rozb4} = 0.530$; $\lambda_{rozb5} = 0.460$; $\lambda_{rozb6} = 0.020$. All items are thus significantly associated with their construct.

The variance explained for the factor is VE = 0.402, and the composite reliability is CR = 0.764. According to Aranowska's formula (cf. formula (2.3), the reliability coefficient is 0.560. A considerable difference between the reliability coefficients is noticeable. The CR coefficient exhibits a clearly higher value, which—as shown earlier—may result from summing correlations and then squaring them, potentially inflating the actual reliability value.

The experienced parental difficulty was operationalized using the *Experienced*

Parental Difficulty Scale. This scale allowed for the assessment of the parent's subjective feeling that raising a child is difficult, frustrating, or emotionally burdensome. This construct does not refer to the child's objective behavior, but rather to the psychological cost of daily functioning in the parental role. The factor loadings (lambdas) of the observed variables are as follows: $\lambda_{tr1} = 0.900$; $\lambda_{tr2} = 0.890$; $\lambda_{tr3} = 0.920$; $\lambda_{tr4} = 0.910$; $\lambda_{tr5} = 0.920$; $\lambda_{tr6} = 0.900$; $\lambda_{tr7} = 0.910$; $\lambda_{tr8} = 0.690$. The explained variance for the factor is VE = 0.780, and the reliability is CR = 0.966. According to Aranowska's formula (see formula (2.3)), the reliability value is 0.883.

The child representation in the parent's mind was reflected through scores obtained on the *Child Representation Scale*. This scale enabled capturing the way in which the parent perceives their child in daily interactions, both cognitively and emotionally. The factor loadings are as follows: $\lambda_{r1} = 0.383$; $\lambda_{r2} = 0.379$; $\lambda_{r3} = 0.816$; $\lambda_{r4} = 0.505$; $\lambda_{r5} = 0.876$; $\lambda_{r6} = 0.828$; $\lambda_{r7} = 0.619$; $\lambda_{r8} = 0.769$. The explained variance for the factor is VE = 0.454, and the reliability is CR = 0.859. According to Aranowska's formula (see formula (2.3)), the reliability value is 0.674.

In summary, the reliability coefficients calculated using Aranowska's formula are consistently lower than the reliability estimated through CR. Therefore, the latent variables are characterized by lower-than-expected reliability and moderate explained variance, which indicates a barely acceptable construction of the measurement model. Subsequently, the relationships between latent variables were compared, and the discriminant power of the model was evaluated. The correlation results between pairs of constructs are presented in Table 2.1, and these values are also illustrated in the graph showing the results of the CFA (Figure 2.2).

Table 2.1. Statistically significant correlations between pairs of constructs calculated based on the relationships estimated in the Confirmatory Factor Analysis

Structures	Discrepancy	Difficulty	Representation
Discrepancy	1.000	0.580	0.290
Difficulty		1.000	0.460
Representation			1.000

All relationships presented in Table 2.1 are statistically significant. Two relationships proposed in the substantive model are marked in red. All remaining associations, which are not estimated by the structural equation model (e.g., the relationship of 0.290), are marked in black. This relationship is treated as theoretically irrelevant in the model and will therefore not be included in the estimation procedure[2].

The relationships between the latent variables do not exceed the lambda values

[2] This relationship was treated as resulting from the co-correlation of variables and thus regarded as a spurious association. It is expected that the discrepancies between the matrices of the saturated model (in which all relationships, including those between the epsilons, are included) and the estimated model (in which only theory-predicted relationships are retained) will be minimal. Verifying whether the saturated and the estimated matrices do not differ significantly constitutes the implementation of the core idea of modeling.

and are lower than the values of explained variance, which demonstrates sufficient discriminant power of the latent variables. Moreover, no cross-loadings of observable variables occurred (each belongs to only one latent variable), which confirms the correct construction of the model (Hair et al., 2006).

In the next step, model fit statistics were calculated and are presented in Table 2.2.

Table 2.2. Fit statistics of the measurement model in the Confirmatory Factor Analysis based on the data

$\chi^2(206) = 1606.333; p < 0.0005$
$df = 207$
$\chi^2/df = 7.760$
CFI = 0.769
RMSEA = 0.146

The model under discussion has a substantial number of degrees of freedom (206), which indicates its high level of complexity resulting from the number of variables included in the analysis[3]. However, the value of the absolute model fit index (RMSEA < 0.08) indicates poor model fit to the data (Hair et al., 2006; Konarski, 2009). The high value of the χ^2 statistic also suggests poor fit, although it should be remembered that this value increases with the number of degrees of freedom. Therefore, the Chi² statistic divided by the number of degrees of freedom is often used. In this case, the value of χ^2/df is 7.760, which further confirms the poor fit of the CFA model to the data.

Additionally, a CFI value below 0.9 indicates interdependence between the latent variables. This statistic assumes independence of the latent variables, which is difficult to achieve when the model is used to verify phenomena of a causal nature. In such cases, the variables remain mutually dependent and determined, as is observed in the model being verified. Despite good reliability and sufficient explained variance, the model proves to be poorly constructed, which suggests that further calculations should not be continued.

The RMSEA statistic clearly indicates the model's misspecification. Releasing one more degree of freedom (by excluding the relationship between discrepancy and the child's representation) would not improve this result. In such a case, the analysis should be discontinued. It could, of course, be continued, but further steps

[3] A large number of degrees of freedom in the CFA model results from the fact that it does not estimate relationships between errors, known as epsilons (of which there are 28), assigned to each observed variable. In contrast, the saturated model includes all possible relationships, including those between epsilons, which leads to a zero number of degrees of freedom ($df = 0$). Since the CFA model includes latent variables, and each of them is associated with its own observed indicators, the number of epsilons—and thus the number of degrees of freedom—increases. In CFA, correlating errors is not allowed, although in practice this assumption is often violated. It is assumed that the unexplained variances of individual items should not be correlated, because such correlation would suggest the existence of an additional, unmodeled variable unknown to the researcher, which affects the results.

will not improve the quality of the model fit, as will be confirmed later in the analysis.

There may be two reasons for this outcome:

1. The theoretical model is incorrect.
2. The measurement model has been poorly constructed.

The researcher must consider both possibilities. In our case, it is known that the assumptions of the measurement model were not fulfilled and the proper nature of the latent structures was not reconstructed. Instead of building a hierarchical measurement model, a single-level model was constructed, despite the fact that our structures contain two factors. As a consequence, the lambda values (factor loadings) fell below 0.7. This violation of the assumption concerning lambda values burdened the model, leading to its poor fit[4].

This error was made deliberately and consciously in order to demonstrate what a poorly fitted model looks like and what consequences arise from neglecting theoretical assumptions when constructing measurement models.

The next chapter will present a measurement model that accounts for the complexity of the latent structures, namely the so-called hierarchical model.

[4] As can be seen, SEM models are extremely strict when it comes to violations of assumptions. A researcher who fails to adhere to these assumptions cannot expect that "something will work out", because the results will be unreliable, and the model is highly likely to be rejected. Even if an SEM model initially appears to fit the data, over time it will fail to fulfill its predictive functions, ultimately leading to its rejection by more advanced methods such as artificial neural networks, as will be discussed in detail later in this work.

CHAPTER 3

Second-Order Hierarchical Measurement Models

3.1. Definition and Theoretical Foundations

In the previous chapters, single-level measurement models were discussed, in which all variables are situated at the same level of analysis. However, empirical reality often requires the consideration of more complex theoretical structures in which variables are arranged hierarchically across different levels. Such structures cannot be fully captured using single-level measurement models.

Hierarchical second-order measurement models address the need to model more complex relationships, in which higher-order variables serve an integrative function in relation to lower-order variables. This allows for the multidimensionality of the analysed constructs to be captured. Such data organisation enables the analysis of relationships both within each level and between levels, which opens new interpretive and analytical possibilities. The hierarchical organisation of constructs within the investigated theory necessitates the use of hierarchical second-order measurement models, since classic first-order models are incapable of capturing inter-level factor relationships or the full structure of the phenomena under analysis.

This chapter presents the theoretical foundations and practical applications of hierarchical second-order measurement models. It discusses the key principles that determine their construction, as well as the consequences of analysing data based on multilevel structures. Two types of hierarchy should be distinguished: structural scope hierarchy and structural element hierarchy. Although these two concepts are related, they refer to different aspects of modeling and require separate discussion.

The structural scope hierarchy refers to situations in which model elements occur at different hierarchical levels. In such structures, relationships between elements may occur both within the same level and across different levels. For example, in a family (as a hierarchical system), parents and children form separate subsystems that may mutually influence one another (Barbaro, 1999). In such modeling structures, it is essential to account for the relationships between elements at different levels, which requires the use of multilevel structural equation models (MSEM) (Garson, 2013; Heck & Thomas, 2009). In these models, elements from higher levels may modify or moderate the relationships at lower levels, which reflects the complexity of the phenomena under investigation (Twisk, Jos, 2010).

In turn, the structural element hierarchy refers to the internal complexity of individual variables or elements within the model. In this context, a hierarchical element consists of several sublevels, with each sublevel containing sub-elements that together constitute the higher-level element. For example, general life satisfaction (a higher-level construct) can be divided into three simpler components: *satisfaction with interpersonal relationships, satisfaction with work, and satisfaction with health*. Each of these components is measured by further, even simpler questionnaire items (e.g., "Are you satisfied with your relationships with close others?", "Does your work give you a sense of fulfillment?", "How would you rate your health status?"). This type of hierarchy is typical for measurement models in which a higher-order latent variable encompasses several lower-order latent variables, which in turn explain observable variables. In practice, this means that higher-level variables are superordinate to lower-level variables, and their mutual relationships and dominance determine how the structure of the model reflects the complexity of the measured construct (Szymańska, 2017a).

In summary, the structural scope hierarchy refers to the complexity of the entire theoretical structure and the relationships between its levels, whereas the structural element hierarchy concerns the internal complexity of individual elements in the model. Understanding these two aspects of hierarchy is crucial for the correct construction and interpretation of multilevel and hierarchical models in the context of structural equations.

Figure 3.1 presents a schematic representation of a multilevel structure in which the elements at the second level (denoted as Y_j) are related to the elements at the first level (denoted as X_j). This structure illustrates the structural scope hierarchy, where relationships between elements occur both within the same level and across different hierarchical levels.

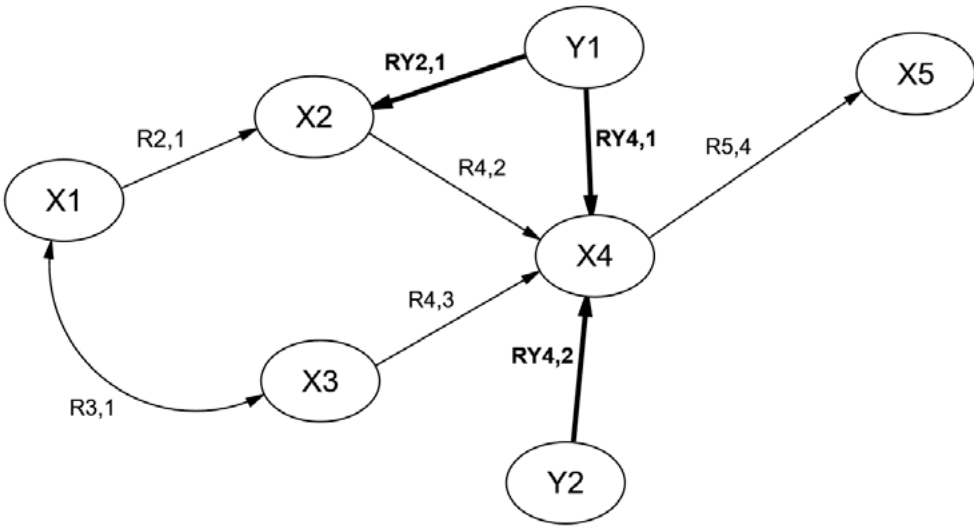


Figure 3.1. Conceptual diagram of a multilevel structure illustrating the structural scope hierarchy

The diagram does not present a content-based example but rather the general structure of relationships between levels in multilevel models. The meaning of the variables (X_j, Y_j) and the interpretation of the relationships between them depend on the operationalization adopted in the study. The figure serves to illustrate the mechanism of hierarchical data organisation within the SEM framework.

The arrows in the diagram symbolise the direction of influence that elements at a higher level exert on elements at a lower level. For example, the element Y_1 may modify the relationship $R_{4,2}$ between two variables from the first level, which indicates its dominant role within the structure. Simultaneously, the element Y_2 may directly affect variable X_4, demonstrating that inter-level relationships are not only complex but also diverse in the nature of their interactions.

It is important to emphasise that hierarchy in this context refers to the complexity of the structure as a whole, where each level adds an additional layer of dependencies and relationships. This distinguishes structural hierarchy from the hierarchy of elements, where internal complexity pertains to individual variables or elements in the model. Figure 3.1 clearly illustrates how elements from different levels may interact with one another, forming a complex network of relationships that reflects the intricacy of the phenomena being studied.

Such hierarchical structures are verified using multilevel structural equation models (MSEM), which enable the analysis of complex relationships between different levels of the structure. MSEM models allow for the inclusion of both within-level and between-level dependencies, which is essential for the correct interpretation of multidimensional phenomena. As emphasised by Heck and Thomas (2009), the application of MSEM models is necessary to capture the full complexity of

hierarchical theoretical structures, which cannot be properly analysed using simpler, single-level structural equation models.

In turn, the structural element hierarchy refers to the internal organisation of a single element within the model. That is, one element may consist of several subordinate levels that together form a more complex whole. Figure 3.2 presents a hierarchical element of structure X_j, which consists of three levels: the lowest level M_j, the intermediate level L_j, and the highest level X_j. In this form of hierarchy, elements from higher levels (e.g., X_j) are superordinate to those from lower levels (e.g., L_j and M_j), meaning that the lower-level elements are constituent parts of the higher-level element and together form an integrated whole (Szymańska, 2017a).

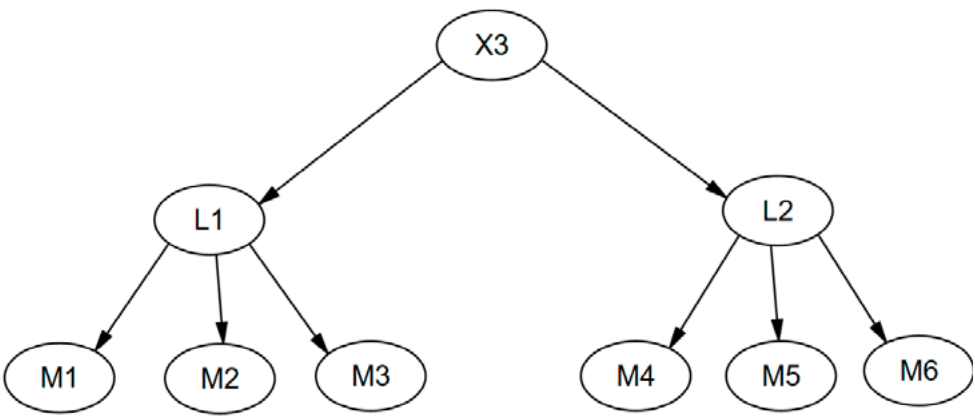


Figure 3.2. Example of a hierarchical structural element illustrating superordinate and subordinate relationships

The diagram presents the general structure of a hierarchical element within a multilevel model. The higher-order variable (X_3) integrates two intermediate factors (L_1 and L_2), each of which explains a set of indicator variables (M_1–M_6). The figure is conceptual and does not depict a specific theoretical or empirical example. Its purpose is to illustrate the logical structure of superordinate and subordinate relationships in hierarchical models.

Figure 3.3 illustrates a complex structure combining both structural scope hierarchy and structural element hierarchy. This diagram merges the characteristics of the two previous figures, simultaneously showing the relationships between levels and the internal complexity of one of the structural elements.

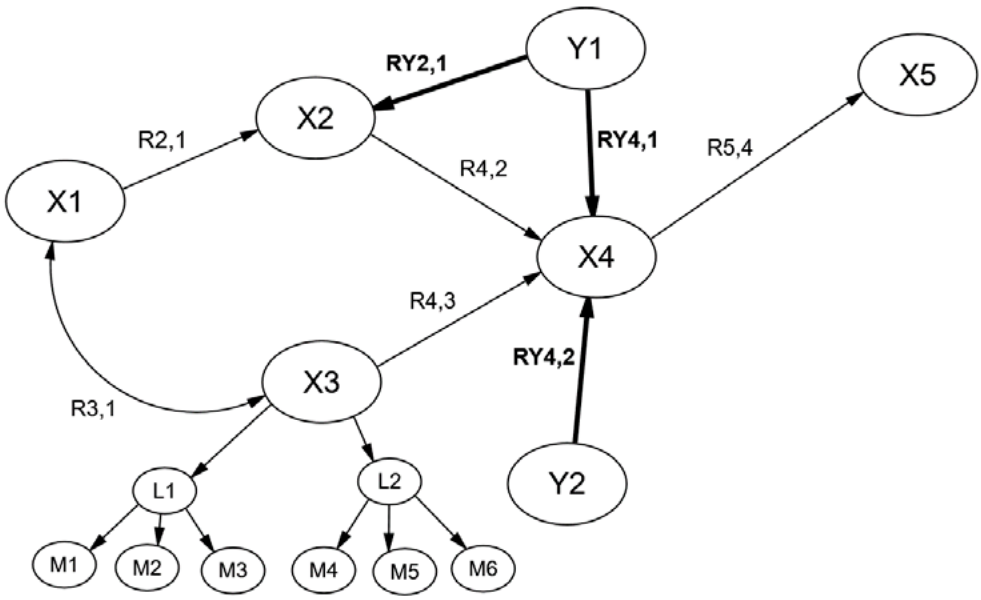


Figure 3.3. Complex multilevel structure integrating structural scope hierarchy and structural element hierarchy

The diagram presents a conceptual example of a complex hierarchical structure in which both inter-level relationships (structural scope hierarchy) and the internal complexity of individual elements (structural element hierarchy) are simultaneously present. The variable X_3 represents an example of a hierarchical variable encompassing the factors L_1 and L_2, which in turn explain sets of indicator variables (M_1–M_6). Simultaneously, the variables Y_j (Y_1, Y_2) influence the relationships between lower-level variables (X_j), illustrating the mechanism of moderation in multilevel models. The figure is illustrative and does not represent a specific theoretical or empirical model.

In the outer layer of the structure, the second-level elements, denoted as Y_j, affect the first-level elements, denoted as X_j, illustrating the structural scope hierarchy. Within one of these elements, marked as X_3, an additional hierarchy is shown, where the second level consists of variables L_j and the first level consists of variables M_j. This configuration allows for the simultaneous understanding of both aspects of hierarchy: the dominance of higher structural levels over lower ones, and the internal complexity of individual elements.

Underestimating structural complexity in modeling may lead to the rejection of a properly constructed model, thereby increasing the risk of committing a Type I error. An example of this issue was presented in Chapter 8.3, where a measurement model was rejected due to the omission of structural element hierarchy. In psychology, the issue of hierarchy plays a particularly important role, as many theories are based on the concept of elements with complex, multilevel structures. Examples

include theories of personality or intelligence, where the complexity of the construct is crucial for its proper understanding.

This chapter presented the conceptual and methodological foundations of second-order hierarchical measurement models, focusing on the logical structure of relationships between levels and the internal organisation of variables. The diagrams used are conceptual in nature and serve to illustrate abstract mechanisms of data organisation within the SEM framework. Empirical examples of the application of hierarchical models, referring to specific psychological constructs, will be presented in later sections of the work, in the context of data analysis and model operationalisation. This approach maintains a distinction between structural analysis and theoretical interpretation, which is consistent with the principles of quantitative methodology.

However, taking this complexity into account requires conducting research on larger samples, which often poses logistical and financial challenges. One possible solution to this problem is to construct models that reduce complexity by focusing on meta-traits. This approach allows for the preservation of the theoretical structure's integrity while reducing the demands on sample size, making the research more feasible in practice (Szymańska, 2017a).

3.2. Formal Assumptions of Hierarchical Measurement Models

In a hierarchical measurement model, the structure of a higher-order latent variable includes lower-order latent variables, which explain the variance of their respective observable variables. An example of such a model is presented in Figure 3.4. In this model, the variable ξ_1 is a higher-order variable composed of two lower-order latent variables, ξ_4 and ξ_5, which explain the variance of their corresponding observable variables (X_1–X_3 and X_4–X_6). The variable ξ_2 is a unidimensional variable that explains the variance of its five observable variables (X_7–X_{14}) (Szymańska, 2017a).

The variable ξ_3, as a higher-order variable, explains the variance of the latent variables ξ_6 and ξ_7. Consequently, ξ_6 and ξ_7 explain the variance of their respective observable variables, namely X_{15}–X_{17} and X_{18}–X_{22}. In such a model, each level of the hierarchical structure is connected to the corresponding observable variables, which allows for a precise analysis of the complexity of the measured construct.

In a hierarchical measurement model, as in a single-level model, observable variables are denoted by the symbol X, and their error variances by the symbol σ (sigma). Factor loadings are denoted by λ (lambda), latent variables by ξ (xi), and residual (structural) errors of latent variables by ζ (zeta). Correlations between latent variables are represented by φ (phi) (Szymańska, 2017a).

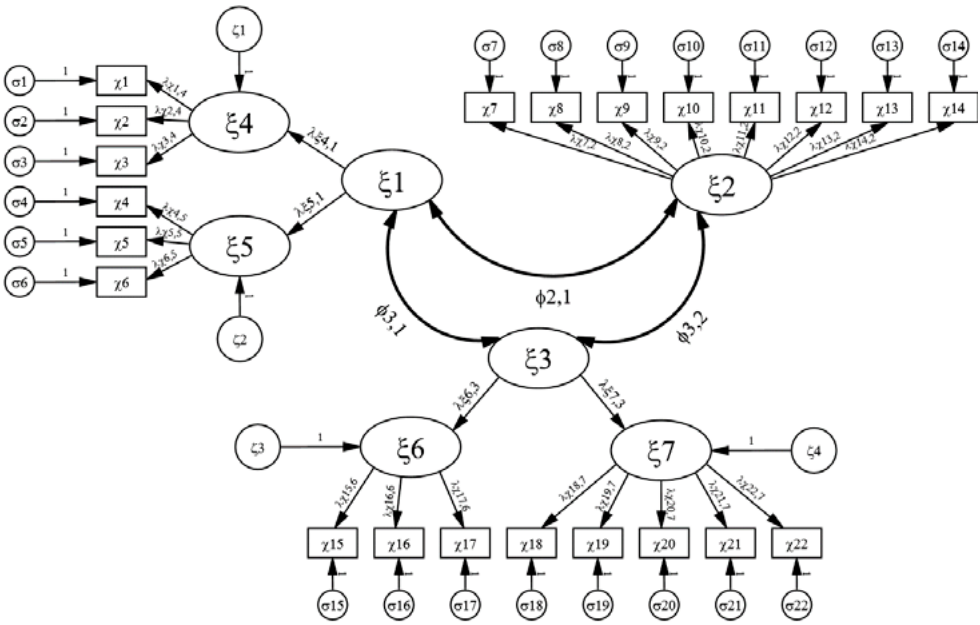


Figure 3.4. Diagram illustrating a hierarchical measurement model, where λ represents the factor loading, σ indicates the error variance of observable variables, χ denotes observable variables, ξ refers to latent variables, ζ (zeta) indicates the residual (structural) error of latent variables, and φ represents correlations between constructs

The figure serves an illustrative purpose—it does not refer to a specific empirical context but solely visualises the formal dependencies within the model and demonstrates how the mathematical notation (matrices λ, φ, ζ, σ) corresponds to the structure of the graphical model. This visualisation facilitates the understanding of subsequent formal notations and the interpretation of model equations.

Verification of the hierarchical structure of the measurement model using Confirmatory Factor Analysis (CFA) follows a procedure similar to that used for a single-level model, as described in detail in Chapter 2.3. In the case of a hierarchical measurement model, this process involves evaluating the fit of the theoretical model to empirical data. However, due to the structural complexity involving multiple levels of latent variables, verification requires additional steps and broader analysis.

CFA enables the testing of hypotheses concerning factor structure and the relationships between latent and observable variables. In the case of a hierarchical model, this entails the analysis of both the relationships at the level of lower-order latent variables and their links to higher-order latent variables. A key element of verification—as in a single-level model—is assessing model fit, which is conducted using indices such as RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), TLI (Tucker–Lewis Index), and χ^2 (Chi-Square).

As in the single-level model, an important step is the analysis of residuals—that is, the differences between the observed and predicted covariance matrices. The residuals should be low and evenly distributed, which indicates proper model fit. In a hierarchical measurement model, particular attention must also be paid to the evaluation of factor loadings at both the level of lower-order latent variables and the higher level, where higher-order variables explain the variance of the lower levels.

Applying these criteria within the hierarchical structure of the measurement model enables a more comprehensive analysis of complex theoretical phenomena, ensuring the accuracy and reliability of the results.

3.3. Application Example of a Hierarchical Measurement Model in Research

To illustrate the significance of incorporating the hierarchical nature of structural elements in modeling, this chapter presents a revised measurement model previously introduced in Chapter 2.3. The introduced modification acknowledges that latent variables such as "discrepancy" and "representation" are not unidimensional constructs but complex, multidimensional structures. This reconstruction of the model allows for a more accurate reflection of the research reality and enhances both the validity and reliability of the obtained results.

A measurement model is essential for determining whether latent structures have been properly constructed. Based on a psychological construct, a hierarchical measurement model was developed, consisting of three latent variables: (a) discrepancy (see "discrepancy" in Figure 3.5), (b) experienced parental difficulty (see "difficulty" in Figure 3.5), and (c) the representation of the child (see "representation" in Figure 3.5). A graph presenting the results of the hierarchical Confirmatory Factor Analysis is shown in Figure 3.5.

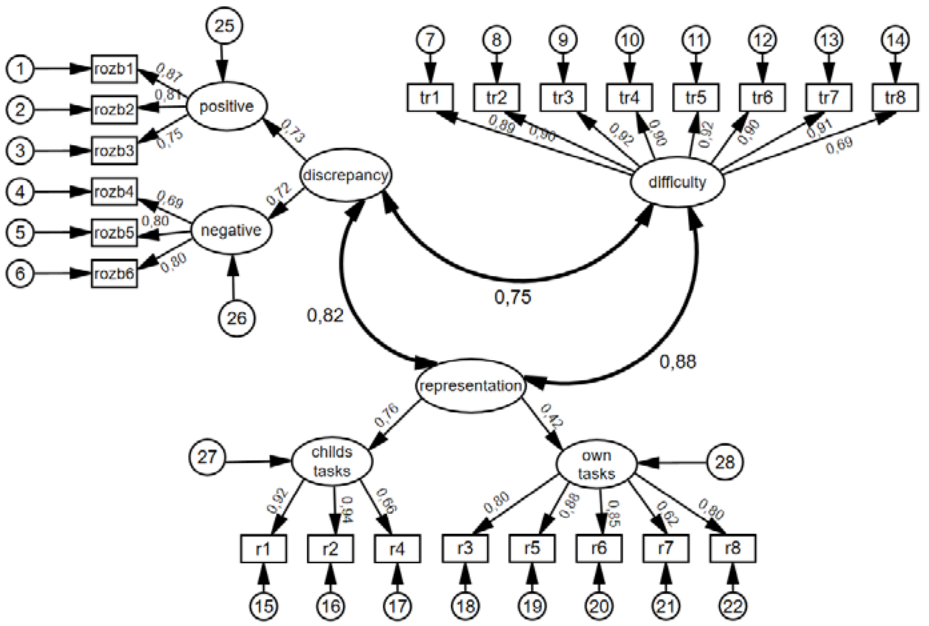


Figure 3.5. Graph presenting the results of the hierarchical Confirmatory Factor Analysis of the discrepancy, difficulty, and child representation model

Discrepancy is a construct measured using the Discrepancy Scale. The lambdas (factor loadings) for the observable variables are as follows: $\lambda_{rozb1} = 0.870$; $\lambda_{rozb2} = 0.810$; $\lambda_{rozb3} = 0.750$; $\lambda_{rozb4} = 0.690$; $\lambda_{rozb5} = 0.800$; $\lambda_{rozb6} = 0.800$. All test items are significantly associated with their respective construct at a statistically significant level. The "discrepancy" construct consists of two factors: the first factor includes questions regarding positive goals (i.e., traits the parent would like to develop in their child), and it includes the observable variables $rozb_1$, $rozb_2$, and $rozb_3$. The lambda (factor loading) for this factor is $\lambda_{positive} = 0.730$. The second factor includes questions regarding negative goals (i.e., traits the parent would not want to develop in their child), and it includes the observable variables $rozb_4$, $rozb_5$, and $rozb_6$. The lambda for this factor is $\lambda_{negative} = 0.720$.

The explained variance for the first factor (labelled "positive" in Figure 3.5) is VE = 0.658, and the reliability CR = 0.852, while the reliability based on Aranowska's γ coefficient is 0.811. The explained variance for the second factor (cf. "negative", Figure 3.5) is similarly VE = 0.585 (reliability CR = 0.808, and reliability based on γ – Aranowska's coefficient is 0.765). For the higher-order structure ("discrepancy"), the explained variance based on these two factors is VE = 0.526 (reliability CR = 0.689, and reliability based on Aranowska's γ coefficient is 0.725). It should be noted that the introduction of the hierarchical structure improved the parameters of the latent structure (lambda values, reliability, and the percentage of explained

variance). Table 3.1 presents a comparison of the structure parameters in the single-level and hierarchical models.

Table 3.1. Comparison of values in the single-level and hierarchical models for the discrepancy structure

Single-Level Model			Hierarchical Model			
Position	**Factor Loadings λ**	**Reliability (CR) Percentage of Explained Variance (VE)**	**Factor Loadings λ**	**Hierarchical Factor Loadings (λ)**	**Reliability (CR) Percentage of Explained Variance (VE)**	
rozb1	0.830	CR = 0.764 VE = 0.402 γ = 0.560	0.870	0.730	CR= 0.852 VE = 0.658 γ = 0.811	CR= 0.689 VE = 0.526 γ = 0.725
rozb2	0.800		0.810			
rozb3	0.770		0.750			
rozb4	0.530		0.690	0.720	CR= 0.808 VE = 0.585 γ = 0.765	
rozb5	0.460		0.800			
rozb6	0.020		0.800			

The obtained results indicate that the construct of *discrepancy* is not a unidimensional structure but consists of two distinct components corresponding to different parental upbringing orientations. The first relates to the discrepancy concerning *positive parental goals*—i.e., traits parents would like to develop in their child (e.g., responsibility, empathy). The second component reflects the discrepancy concerning *negative parental goals*—i.e., undesirable traits that parents seek to avoid developing in their child (e.g., aggressiveness, selfishness). This distinction is theoretically important, as it shows that parents may experience discrepancy both when the child fails to meet their positive expectations and when the child displays traits they want to avoid. The hierarchical model thus makes it possible to capture the complexity and directionality of parental expectations, which aligns with cognitive and motivational assumptions regarding the regulation of upbringing goals.

Experienced upbringing difficulty is a variable measured using the Experienced Upbringing Difficulty Scale. The lambdas (factor loadings) for the observable variables (eight test items) are as follows: $\lambda_{tr1} = 0.900$, $\lambda_{tr2} = 0.890$, $\lambda_{tr3} = 0.920$, $\lambda_{tr4} = 0.900$, $\lambda_{tr5} = 0.920$, $\lambda_{tr6} = 0.900$, $\lambda_{tr7} = 0.913$, $\lambda_{tr8} = 0.694$. All these items are statistically significantly associated with the construct they represent. The explained variance for this factor is VE = 0.779, and the reliability CR = 0.965, while the reliability based on Aranowska's γ coefficient was 0.882. The high level of reliability and the good explained variance indicate that the latent variable *experienced upbringing difficulty* is very well constructed. Since this structure remained single-level, its parameters did not change significantly compared to other models analysed.

The obtained results confirm that *experienced upbringing difficulty* has a unidimensional structure and very high factor loadings. This means that the construct constitutes an internally coherent experience, strongly linked to the overall level of parental tension and burden.

Child representation is another variable measured using the Child Representation in the Parent's Mind Scale. The lambdas (factor loadings) for the observable variables (eight test items) are as follows: $\lambda_{r1} = 0.920$; $\lambda_{r2} = 0.940$; $\lambda_{r3} = 0.800$; $\lambda_{r4} = 0.660$; $\lambda_{r5} = 0.880$; $\lambda_{r6} = 0.850$; $\lambda_{r7} = 0.620$; $\lambda_{r8} = 0.800$. All these items are statistically significant with respect to their construct. This construct includes two factors. The first factor concerns the representation of the child's tasks as less important compared to the parent's tasks (labelled in Figure 3.5 as "child's tasks"). The lambda (factor loading) for this factor is $\lambda_{childtasks} = 0.760$. The second factor refers to the representation of the parent's tasks as more important than the child's tasks (labelled in Figure 3.5 as "one's own tasks"). The lambda for this factor is $\lambda_{owntasks} = 0.420$.

The explained variance for the first factor is VE = 0.722, and the reliability CR = 0.884; according to Aranowska's gamma coefficient, the reliability was 0.850, indicating very high reliability and good explained variance of this variable. The explained variance for the second factor is slightly lower: VE = 0.632, CR = 0.895, and according to Aranowska's gamma coefficient, the reliability was 0.795, which also indicates very good reliability and good explained variance. For the higher-order structure, i.e. representation, the explained variance based on these two factors is VE = 0.377, CR = 0.527, and 0.614 according to Aranowska's coefficient. Similarly to the discrepancy structure, the representation structure also shows changed values, which are presented in Table 3.2.

Table 3.2. Comparison of values in the single-level and hierarchical models for the representation structure

Single-Level Model			Hierarchical Model			
Position	**Factor Loadings λ**	**Reliability (CR) Percentage of Explained Variance (VE)**	**Factor Loadings λ**	**Hierarchical Factor Loadings (λ)**	**Reliability (CR) Percentage of Explained Variance (VE)**	
r1	0.383	CR= 0.859 VE = 0.454 γ = 0.674	0.920	0.760	CR= 0.884 VE = 0.722 γ = 0.850	CR= 0.527 VE = 0.377 γ = 0.614
r2	0.379		0.940			
r4	0.816		0.660			
r3	0.505		0.800	0.420	CR= 0.895 VE = 0.632 γ = 0.795	
r5	0.876		0.880			
r6	0.828		0.850			
r7	0.619		0.620			
r8	0.769		0.800			

The structure of the variable "child representation" confirmed the theoretical assumptions—two representations were identified: one concerning the parent's own tasks and one concerning the child's tasks. This means that Gurycka's concept was successfully reflected empirically in the construction of the scale.

It can be observed that the constructed latent variables are characterised by high or very high reliability values. The variance explained by these variables also reaches a moderate or high level, indicating good quality of the variable construction. Overall, the measurement model confirmed that the latent variables were correctly built and can serve as a solid foundation for further analyses.

Next, an analysis was conducted to compare the relationships between latent variables and to assess whether the constructed model demonstrates discriminant validity. Based on the relationships estimated in the Confirmatory Factor Analysis (CFA), correlations between pairs of constructs were calculated. The results of these correlations are presented in Table 3.3.

Table 3.3. Statistically significant correlations between pairs of constructs calculated based on the relationships estimated in the Confirmatory Factor Analysis

Structures	Discrepancy	Difficulty	Representation
Discrepancy	1.000	**0.750**	0.820
Difficulty		1.000	**0.880**
Representation			1.000

All presented relationships are statistically significant. As shown in Table 3.3, the correlations between the latent variables do not exceed the lambda values (factor loadings) of these variables. This means that the latent variables demonstrate construct coherence, as they are better explained by their own observable indicators than by other latent variables. In other words, the latent variables correlate more strongly with their own indicators, which confirms the correct construction of the model. The formula for verifying the internal coherence of a latent variable was developed by Aranowska. It will be discussed in later chapters, which will address the use of artificial neural networks in combination with SEM models.

The high correlation between "discrepancy" and "difficulty" indicates that the greater the discrepancy between a parent's upbringing goals (i.e., the traits they wish to develop in their child) and the child's current level of development in those areas, the higher the level of experienced parental stress. The relationship between "difficulty" and "representation" suggests that the more negative the image of the child in the parent's mind, the greater the subjective sense of difficulty in the upbringing process. In turn, the correlation between "discrepancy" and "representation" may indicate that large discrepancies between expectations and reality can influence the way the child is perceived, in line with cognitive theories of interpreting the child's behaviour through the lens of parental goals and emotions.

The next step was to calculate model fit statistics. Table 3.4 presents the fit statistics values for two models: the single-level measurement model and the hierarchical measurement model.

Table 3.4. Fit Statistics Values for the Measurement Models

Single-Level Model	Hierarchical Model
$\chi^2(207) = 1606.333$; $p < 0.0005$	$\chi^2(202) = 678.526$; $p < 0.0005$
$df = 207$	$df = 202$
$\chi^2/df = 7.760$	$\chi^2/df = 3.359$
CFI = 0.769	CFI = 0.921
RMSEA = 0.146	RMSEA = 0.086

The hierarchical model under discussion has 202 degrees of freedom. The lower number of degrees of freedom results from the inclusion of three additional parameters compared to the single-level model, while maintaining the same number of observed variables. As can be seen, all fit indices improved after applying the hierarchical measurement model.

The chi-square (χ^2) value decreased from 1606.333 to 678.526, which is a significant improvement; the χ^2/df ratio decreased from 7.760 to 3.359; and the CFI increased to a value considered acceptable, indicating that the model can be regarded as well-fitted to the data. Moreover, RMSEA decreased from 0.146 to 0.086, which also suggests that the model reached a level close to acceptable model fit[5].

The improvement in fit in the hierarchical model indicates that treating "discrepancy" and "child representation" as complex constructs more accurately reflects the real psychological processes occurring in the parent–child relationship. The hierarchical model, which accounts for the multidimensionality of parental goals and the complexity of the cognitive image of the child, corresponds more closely to the assumptions of Gurycka's theory. According to this theory, parental representations and difficulties are related to the degree of congruence between upbringing goals and the actual developmental level of the child. The results confirm that only a model incorporating these relationships allows for a more complete verification of the theory.

It is worth noting that the introduction of the hierarchical structure led to a significant improvement in model fit. In the following chapter, we will discuss how to compute the system of structural equations based on such a measurement model.

[5] To consider a model well-fitted, the RMSEA value should be RMSEA ≤ 0.08 (Hair et al., 2006); however, some authors suggest that an RMSEA value ≤ 0.10 is also acceptable.

CHAPTER 4

Structural Model – Theory and Applications

4.1. Theoretical Foundations and Formal Assumptions of Structural Models with a Single-Level Measurement Model

Substantive theory refers to the conceptualisation of the relationships between constructs in the analysed model. The structural model, in turn, is a formal representation of that theory, most often presented in the form of a diagram (although a diagram is not essential, it significantly facilitates interpretation). In such a structural model, the relationships between latent variables (constructs) are defined in terms of dependency relations.

Latent variables can be classified as endogenous (dependent) or exogenous (independent), depending on the theory being tested by the model. Endogenous variables are denoted by the symbol η and are associated with an error term denoted as ζ. Exogenous variables, which are independent, are denoted by the symbol ξ and are not defined with an error term. Relationships between endogenous and exogenous variables are represented by the symbol γ, while dependencies between endogenous variables are denoted as β. Observable variables linked to endogenous variables are denoted as Y, with their errors as ε. Observable variables linked to exogenous variables are denoted as X, with their errors as σ (Hair et al., 2006; Szymańska, 2016b). Models defined in this way can be presented in the form of a so-called *path diagram*, as shown in Figure 4.1.

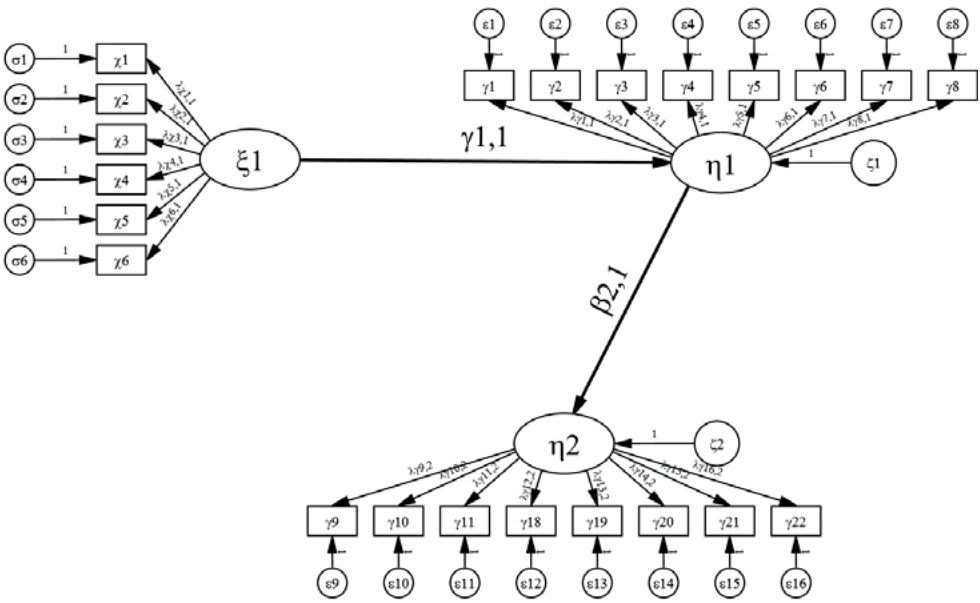


Figure 4.1. Example path diagram for a structural model, where λ represents the factor loading, σ and ε denote the error variances of observable variables, X refers to exogenous observable variables, Y to endogenous observable variables, ξ symbolises the exogenous latent construct, η the endogenous latent construct, γ represents the relationship between the exogenous and endogenous construct, and β the relationship between endogenous constructs

The structural equation model has the following form:

$$B\eta = \Gamma\xi + \zeta,$$

where B is a square coefficient matrix of dimension [m × m], Γ is a rectangular coefficient matrix of dimension [m × n], and ζ is a random vector of residuals (see Aranowska, 2005). These equations can be expressed in detail in matrix form. For the structural model presented in Figure 4.1, the structural equation matrix takes the following form:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_{1,1} & 0 \\ 0 & \beta_{2,1} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \eta_1 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

For the measurement model, the matrices take the form of the following system of equations:

1. For the exogenous variable

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} = \begin{pmatrix} \lambda_{x1,1} \\ \lambda_{x2,1} \\ \lambda_{x3,1} \\ \lambda_{x4,1} \\ \lambda_{x5,1} \\ \lambda_{x6,1} \end{pmatrix} \begin{pmatrix} \xi_1 \end{pmatrix} + \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{pmatrix}$$

2. For the endogenous variables

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \end{pmatrix} = \begin{pmatrix} \lambda_{y1,2} & 0 \\ \lambda_{y2,2} & 0 \\ \lambda_{y3,2} & 0 \\ \lambda_{y4,2} & 0 \\ \lambda_{y5,2} & 0 \\ \lambda_{y6,2} & 0 \\ \lambda_{y7,2} & 0 \\ \lambda_{y8,2} & 0 \\ 0 & \lambda_{y9,3} \\ 0 & \lambda_{y10,3} \\ 0 & \lambda_{y11,3} \\ 0 & \lambda_{y12,3} \\ 0 & \lambda_{y13,3} \\ 0 & \lambda_{y14,3} \\ 0 & \lambda_{y15,3} \\ 0 & \lambda_{y16,3} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{pmatrix}$$

When testing the structural model, two aspects are of key importance: the overall fit of the model to the empirical data (i.e., to the "input matrix"[6]) and the magnitude, direction, and significance of individual parameters. If the model fit is satisfactory and the factor loadings are statistically significant and of appropriate magnitude, the model is considered to be well-fitted.

Model fit can be assessed using various methods, such as ML (*Maximum Likelihood*), GLS (*Generalized Least Squares*), ULS (*Unweighted Least Squares*), Scale-Free Least Squares, or *Asymptotically Distribution-Free*. The ML method is the most popular in the case of continuous variables, as it is the only one capable of handling missing data. However, this method requires the variables to be normally distributed. If the distributions deviate from normality, it is recommended to use alternative verification methods, as described by Konarski (2009). Aranowska (1996) emphasizes that the normality assumption is not always met; therefore, it is advisable to check whether the distributions approximate normality and whether the condition of homoscedasticity is fulfilled.

Model fit means that the resulting matrix of coefficients does not statistically differ from the input matrix. Numerous statistics are available for assessing model fit, which can be classified into three categories: a) *Absolute Fit Measures*, b) *Incremental Fit Indices*, and c) *Parsimony Fit Indices* (Konarski, 2009). The key statistic for evaluating model fit is the Chi-square (Chi^2) test. Its lack of statistical significance indicates good model fit, meaning that there are no significant differences between the resulting matrix and the input matrix. However, due to Chi^2's sensitivity to sample size and model complexity, its value increases with the number of degrees of freedom, which may lead to statistical significance even in well-fitted models (Aranowska, 1996). Therefore, other model fit statistics are commonly used as well.

In the present section, we focus on a detailed presentation of representative examples and formulas for the most commonly used measures from each of the aforementioned categories of model fit assessment methods.

One of the most widely used indicators of model fit is the RMSEA (*Root Mean Square Error of Approximation*). This measure is based on the distribution of the variable in the population, which makes it more accurately reflect the extent to which the model is adequate for the studied population compared to other statistics. RMSEA, as expressed in equation (4.2), is a measure of approximation error, meaning that it represents the level of model misfit; the lower the RMSEA value, the better the model fit. Optimal RMSEA values are <0.01; however, in practice, values usually fall within the range 0.02 < RMSEA < 0.08, which is interpreted as indicating good model fit (Hair et al., 2006; Konarski, 2009; Szymańska, 2016).

$$\text{RMSEA} = \sqrt{\max\left(\frac{Chi^2/(N-1)}{df} - \frac{1}{N-1}\right)}, \tag{4.2}$$

[6] In AMOS software, a model that includes a fully specified input matrix is referred to as a *saturated model*.

where: Chi^2 – is the value of the statistic for the estimated model, N – the size of the examined sample, df – the degrees of freedom of the estimated model.

Incremental fit indices are used to assess the extent to which a given model fits the data better than alternative models, typically compared to the null model. Key indicators in this category include NFI, CFI, TLI, and RNI. Among them, the CFI index is particularly widely used, with values ranging from 0 to 1. A CFI > 0.9 indicates good model fit, while a CFI < 0.9 suggests poor fit. A significant advantage of CFI is its relative robustness to the number of degrees of freedom, which makes it a useful tool for evaluating more complex models, where other statistics may not yield reliable results.

However, CFI assumes the independence of latent structures, which in models verifying causal relationships is an assumption difficult to confirm both theoretically and practically. In such cases, CFI may yield lower values than RMSEA, which, in turn, does not require such an assumption and is therefore often considered a more universal indicator (Szymańska, 2016b; Szymańska & Aranowska, 2016).

$$\text{CFI} = 1 - \frac{Chi^2 - df}{{Chi^2}_n - df_n}, \tag{4.3}$$

where: ${Chi^2}_n$ – is the value of the statistic for the independent model, df_n – the degrees of freedom of the independent model.

Parsimony Fit Indices assess how well a model fits the data in relation to its complexity. They are primarily useful when comparing two models, allowing one to determine which better captures the analysed relationships. Among the most commonly used statistics of this type are PRI, PGFI, and PNFI. PNFI, in particular, is widely applied, and the closer its value is to one, the better the model fit can be assumed, taking into account its complexity. In general, a PNFI value above 0.60 is considered to indicate good model fit, although some researchers accept even lower values, such as 0.50 (Szymańska, 2016b; Szymańska & Aranowska, 2016).

$$\text{PNFI} = \frac{df}{df_n}\ \text{NFI}, \tag{4.4}$$

$$\text{NFI} = \frac{{Chi^2}_n - Chi^2}{{Chi^2}_n}.$$

It is worth noting that better model fit, as assessed using parsimony indices, may result not only from the model's actual fit but also from its simplicity. Simpler models, containing fewer constructs and observable variables, are associated with a smaller number of degrees of freedom, which may lead to more favourable results in fit measures. Models can be compared in various ways, for instance, by testing the same model across different populations, which is recommended. Another approach involves comparing two models within the same population. When the models differ only

in the direction of relationships between constructs and not in complexity, the nested models method is applied. In such cases, two opposing regression paths are specified between the constructs, forming what is referred to as a non-recursive model. This method is used to test alternative models within a single framework. If two models are compared and one is more complex, it is essential to determine whether the *Chi*² difference, given the corresponding difference in degrees of freedom, is statistically significant (Szymańska, 2016b; Szymańska & Aranowska, 2016). For structural models, degrees of freedom are calculated according to formula 4.5:

$$df = (p + q)\,(p + q + 1)/\,2 - t, \tag{4.5}$$

where: p is the number of observable variables for exogenous structures, q is the number of observable variables for endogenous structures, and t is the number of parameters estimated in the model.

Alternative models are compared using Formula 4.6:

$$\Delta\chi^2_{\Delta df} = \chi^2_{df(B)} - \chi^2_{df(A)}$$
$$\Delta df = df(B) - df(A) \tag{4.6}$$

If $\Delta\chi^2$ is statistically significant, it should be concluded that the more parsimonious model (i.e., the one with more degrees of freedom) is the better model.

The efficient use of degrees of freedom is crucial in structural modeling. It is assumed that correlations not directly accounted for in the model are indirectly represented through the relationships included in it. In other words, parsimony in degrees of freedom involves the deliberate omission of certain relationships between structures that are not essential to the theory under investigation. A high model fit accompanied by a large number of degrees of freedom—i.e., a parsimonious model—indicates high validity. Such a model includes only those paths that are essential for fully understanding the analysed phenomenon, eliminating less significant relationships.

4.2. Theoretical Foundations and Formal Assumptions of Structural Models with a Hierarchical Measurement Model

In structural models with a hierarchical measurement structure, only those relationships that directly follow from theoretical assumptions are included. Other associations that are not essential to the verified theory are omitted from the model as relationships not theoretically justified. Similar to models based on single-level components, endogenous variables are denoted by the symbol η, each accompanied by a residual (structural) error (ζ). These represent the difference between the predicted

value of the latent variable (based on the model) and the actual value of that variable. In other words, residual errors indicate the extent to which the model fails to fully account for the variability of the latent variable. Exogenous, or independent, variables are denoted by the symbol ξ and are not associated with error terms (see Figure 4.2). Relationships between endogenous and exogenous variables are represented by the symbol γ, whereas dependencies among the endogenous variables themselves are denoted by β. Observable variables for the endogenous variables are represented by the symbol Y, with their errors denoted by ε; observable variables for the exogenous variables are denoted as X, with their errors marked as σ. Figure 4.2 presents a sample path diagram of a structural model with a hierarchical measurement structure.

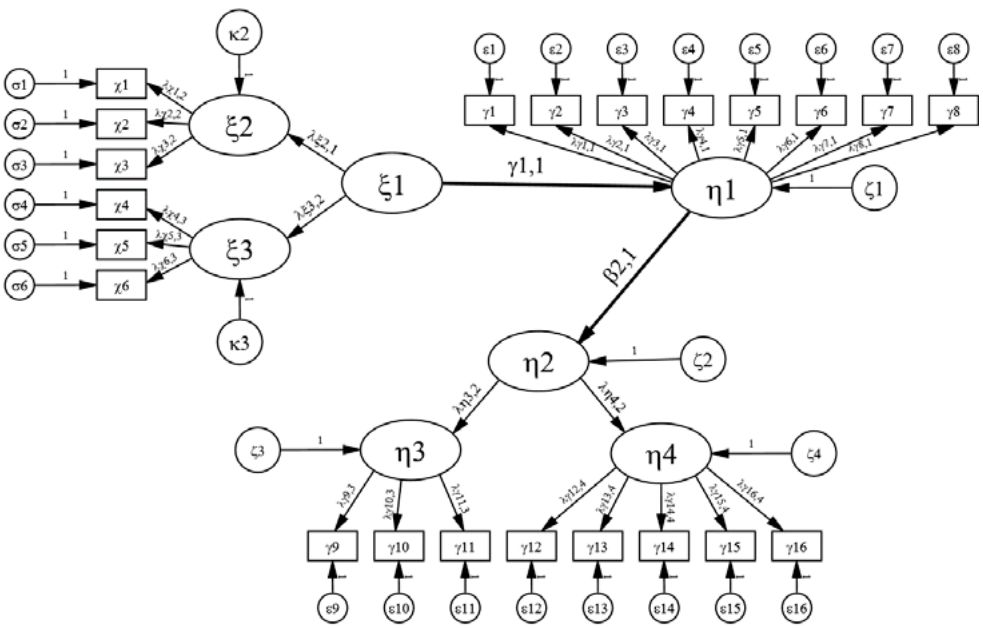


Figure 4.2. Example path diagram for a structural model with a hierarchical measurement structure, in which λ represents factor loadings, σ and ε denote the error variance of observable variables, X indicates exogenous observable variables, Y corresponds to endogenous observable variables, ξ symbolizes the exogenous latent construct, and η reflects the endogenous latent construct. In the diagram, γ denotes the relationship between exogenous and endogenous constructs, β represents the relationships among endogenous constructs, and ζ refers to the residuals of the latent (explained) variables

As can be observed, the variance of the latent structure ξ_1 is determined by the variances of the structures ξ_2 and ξ_3, whereas the variance of the latent structure η_2 depends on the variances of the structures η_3 and η_4. The structures ξ_1 and η_2 are hierarchical in nature, while the structures ξ_2, ξ_3, η_1, η_3, and η_4 are single-level structures.

The matrix of structural equations for the model presented in Figure 4.2 is as follows:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \gamma_{1,1} & 0 \\ 0 & \beta_{2,1} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \eta_1 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix}$$

For the measurement model, the matrices take the form of the following system of equations:

1. For the exogenous variable

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix} = \begin{pmatrix} \lambda_{x1,1} & 0 \\ \lambda_{x2,1} & 0 \\ \lambda_{x3,1} & 0 \\ 0 & \lambda_{x4,1} \\ 0 & \lambda_{x5,1} \\ 0 & \lambda_{x6,1} \end{pmatrix} \begin{pmatrix} \xi_2 \\ \xi_3 \end{pmatrix} + \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{pmatrix}$$

2. For first-level endogenous variables

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{16} \end{pmatrix} = \begin{pmatrix} \lambda_{y1,1} & 0 & 0 \\ \lambda_{y2,1} & 0 & 0 \\ \lambda_{y3,1} & 0 & 0 \\ \lambda_{y4,1} & 0 & 0 \\ \lambda_{y5,1} & 0 & 0 \\ \lambda_{y6,1} & 0 & 0 \\ \lambda_{y7,1} & 0 & 0 \\ \lambda_{y8,1} & 0 & 0 \\ 0 & \lambda_{y9,3} & 0 \\ 0 & \lambda_{y10,3} & 0 \\ 0 & \lambda_{y11,3} & 0 \\ 0 & 0 & \lambda_{y12,4} \\ 0 & 0 & \lambda_{y13,4} \\ 0 & 0 & \lambda_{y14,4} \\ 0 & 0 & \lambda_{y15,4} \\ 0 & 0 & \lambda_{y16,4} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_3 \\ \eta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{pmatrix}$$

For second-level endogenous variables

$$\begin{pmatrix} \eta_3 \\ \eta_4 \end{pmatrix} = \begin{pmatrix} \lambda_{\eta 3,2} & 0 \\ 0 & \lambda_{\eta 4,2} \end{pmatrix} \begin{pmatrix} \eta_2 \end{pmatrix} + \begin{pmatrix} \zeta_3 \\ \zeta_4 \end{pmatrix}$$

4.3. Applications of Structural Models in Psychological and Social Research

This chapter presents applications of structural models in psychological research. It begins with a discussion of a structural system based on a single-level measurement model. Subsequently, a more complex structural model will be presented, based on a hierarchical measurement model that incorporates different levels of latent variables. Such a structure allows for a more accurate representation of complex relationships between variables, which is particularly important in the case of intricate theoretical constructs. These examples will illustrate how different modeling approaches can be applied in research practice, depending on the specificity of the analysed variables and the objectives of the study.

The assumptions concerning the relationships between variables derive directly from the adopted theoretical framework, in which the discrepancy between a parent's parental goal and the child's developmental level constitutes the source of the experienced parental difficulty. This, in turn, affects the parent's representation of the child. The adopted configuration of variables reflects the directional, causal nature of these relationships and is based on the concept of cognitive-emotional determinants of parental mistakes (Gurycka, 1990; Szymańska & Aranowska, 2016).

4.3.1. Structural Model with a Single-Level Measurement Model

The theoretical model described in Section 2.3 of Part I was tested using structural equation modeling (SEM). The verification included only those relationships that had been theoretically specified. This chapter focuses on a detailed analysis of the model.

The presented theoretical model includes three variables: Discrepancy, Experienced Parental Difficulty, and the Representation of the Child in the Parent's Mind. The exogenous variable in the model is Discrepancy, while Experienced Parental Difficulty and the Representation of the Child in the Parent's Mind are endogenous variables. Experienced Parental Difficulty serves as an outcome variable in relation to Discrepancy and simultaneously as a predictor for the Representation of the Child in the Parent's Mind.

Due to the unidirectionality of all paths, this model belongs to the category of recursive models (Gaul & Machowski, 1987). For the purpose of model verification, three scales were developed to measure the analysed traits.

The following research hypotheses were formulated:

H1: The discrepancy between the parent's parental goal and the child's current developmental level in the area of cultivated traits is positively associated with the parental difficulty experienced in upbringing situations.

H2: The parental difficulty experienced in upbringing situations is positively associated with the development of a negative representation of the child and their tasks as represented in the parent's mind.

It should be noted, however, that the model already failed to reflect the theoretical complexity of the constructs at the measurement stage, which led to its rejection. Since it did not meet the criteria of validity and reliability at the measurement level, there is even less reason to expect that it would prove adequate after estimating the full structural model. Nevertheless, for the sake of clarity, an analysis was conducted to illustrate these limitations and help the reader understand why this model was not considered appropriate.

The model was estimated using the *Maximum Likelihood* method. The model fit results are presented in Table 4.1, which includes both the obtained index values and their recommended values—that is, the thresholds each index should reach for the model to be considered well-fitted. For clarity, the *Chi*^{2} statistic values for the independent model are also provided.

Table 4.1. Model Fit Statistics

	Fit Indices	Value	Recommended Value for Failing to Reject H_0	Statistical Significance Level
	χ^2	1606.476	*ns*	$p < 0{,}001$
	df	208		
	n	319		
	χ^2*independent*	6292.969		$p< 0{,}001$
	df independent	231		
Absolute Fit Measures	Hoelther	49		$p = 0{,}05$
Relative Fit Measures – Type 1	NFI	0.745		
	RFI	0.716		
Relative Fit Measures – Type 2	IFI	0.770	>0,900	
Relative Fit Measures – Type 3	CFI	0.769	>0.900	
Fit Measures Accounting for Model Complexity	PNFI	0.671		
	PRNI			
	PCFI	0.693		
	PRATIO	0.900		
Approximation Error Measures	RMSEA	0.145	< 0,06< 0,08	90% probability of model fit

The measures presented in Table 4.1 indicate a lack of fit between the proposed model and the data. In particular, the value of the RMSEA statistic (exceeding the critical threshold of 0.08) and the CFI value (0.900) suggest that the model does not adequately represent the data. In other words, the null hypothesis—assuming no differences between the theoretical and empirical models—should be rejected.

Additionally, as noted by Aranowska (1996), fit statistics in SEM models may sometimes indicate good fit even when the model poorly reflects the actual phenomenon. This especially occurs when the model is not very complex and the sample size is not very large. Such is the case with the present model, which, despite a large sample size, is not particularly complex. It is therefore known with certainty that the model is inadequate, and there is a clear hypothesis as to why. The model fails to reflect the assumed hierarchical structure of the measurement variables—its formal configuration does not correspond to the theoretically postulated arrangement of relationships between variable levels, resulting in a distortion of the construct's representation. Before drawing further conclusions, however, it is worth taking a closer look at the relationships between the variables.

Figure 4.3 presents the path diagram of the estimated model, the structural part of which is represented by the following system of equations:

$$\begin{pmatrix} \text{difficulty} \\ \text{representation} \end{pmatrix} = \begin{pmatrix} \gamma_{\text{difficulty, discrepancy}} & 0 \\ 0 & \beta_{\text{representation, difficulty}} \end{pmatrix} \begin{pmatrix} \text{discrepancy} \\ \text{difficulty} \end{pmatrix} + \begin{pmatrix} \zeta_{23} \\ \zeta_{24} \end{pmatrix}$$

For the measurement model, the matrices take the form of the following system of equations:

1. For the exogenous variable

$$\begin{pmatrix} \text{rozb1} \\ \text{rozb2} \\ \text{rozb3} \\ \text{rozb4} \\ \text{rozb5} \\ \text{rozb6} \end{pmatrix} = \begin{pmatrix} \lambda_{\text{rozb1, discrepancy}} \\ \lambda_{\text{rozb2, discrepancy}} \\ \lambda_{\text{rozb3, discrepancy}} \\ \lambda_{\text{rozb4, discrepancy}} \\ \lambda_{\text{rozb5, discrepancy}} \\ \lambda_{\text{rozb6, discrepancy}} \end{pmatrix} \begin{pmatrix} \text{discrepancy} \end{pmatrix} + \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{pmatrix}$$

1. For first-level endogenous variables

$$\begin{pmatrix} tr1 \\ tr\ 2 \\ tr\ 3 \\ tr\ 4 \\ tr\ 5 \\ tr\ 6 \\ tr\ 7 \\ tr\ 8 \\ r1 \\ r2 \\ r3 \\ r4 \\ r5 \\ r6 \\ r7 \\ r8 \end{pmatrix} = \begin{pmatrix} \lambda_{tr1,\,difficulty} & 0 \\ \lambda_{tr2,\,difficulty} & 0 \\ \lambda_{tr3,\,difficulty} & 0 \\ \lambda_{tr4,\,difficulty} & 0 \\ \lambda_{tr5,\,difficulty} & 0 \\ \lambda_{tr6,\,difficulty} & 0 \\ \lambda_{tr7,\,difficulty} & 0 \\ \lambda_{tr8,\,difficulty} & 0 \\ 0 & \lambda_{r1,representation} \\ 0 & \lambda_{r2,representation} \\ 0 & \lambda_{r3,\,representation} \\ 0 & \lambda_{r4,\,representation} \\ 0 & \lambda_{r5,\,representation} \\ 0 & \lambda_{r6,\,representation} \\ 0 & \lambda_{r7,\,representation} \\ 0 & \lambda_{r8,\,representation} \end{pmatrix} \begin{pmatrix} difficulty \\ representation \end{pmatrix} + \begin{pmatrix} \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \\ \varepsilon_{20} \\ \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix}$$

Figure 4.3 presents the standardized results, while Figure 4.4 shows the unstandardized ones. The first hypothesis assumed that the discrepancy between the parent's parental goal and the gap between the child's current developmental level and that goal is positively related to the parental difficulty experienced. The relationship between the Discrepancy construct and Experienced Parental Difficulty proved to be statistically significant, $\gamma_{11} = 0.58$; $p < 0.0005$. Thus, the first hypothesis could be confirmed (if the model were valid)—it was demonstrated that Discrepancy significantly and strongly correlates with the parental difficulty experienced. Discrepancy accounts for 33.6% of the variance in Experienced Parental Difficulty, as calculated by: $(0.58)^2 = 0.336$.

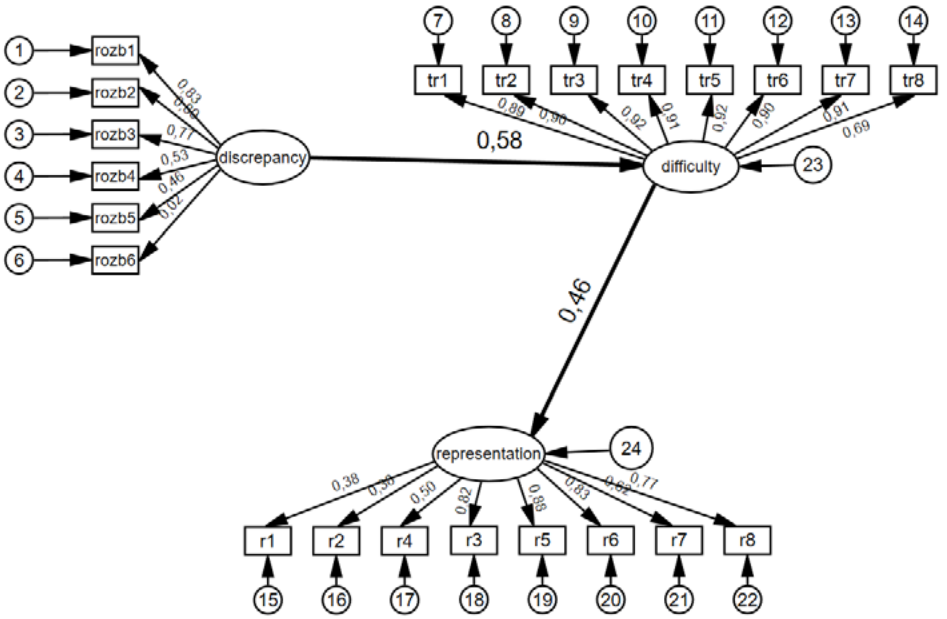


Figure 4.3. Graphical representation of the model constructed using structural equations. Standardized results.

The regression coefficient for this relationship is 1.33, which means that each one-unit increase in Discrepancy leads to a 1.33-unit increase in the parental difficulty experienced (see Figure 4.4).

The second hypothesis assumed that Experienced Parental Difficulty is positively correlated with the Representation of the Child in the Parent's Mind. This relationship also proved to be statistically significant, $\beta_{21} = 0.46$; $p < 0.0005$. This result would confirm the validity of the second hypothesis—it was demonstrated that the relationship between these constructs is positive and moderate. Experienced Parental Difficulty explains 21% of the variance in the child's representation, $(0.46)^2 = 0.21$. The standardized coefficient for this relationship is 0.18, meaning that a one-unit increase in difficulty results in a 0.18-unit increase in the discussed representation of the child in the parent's mind (see Figure 4.4).

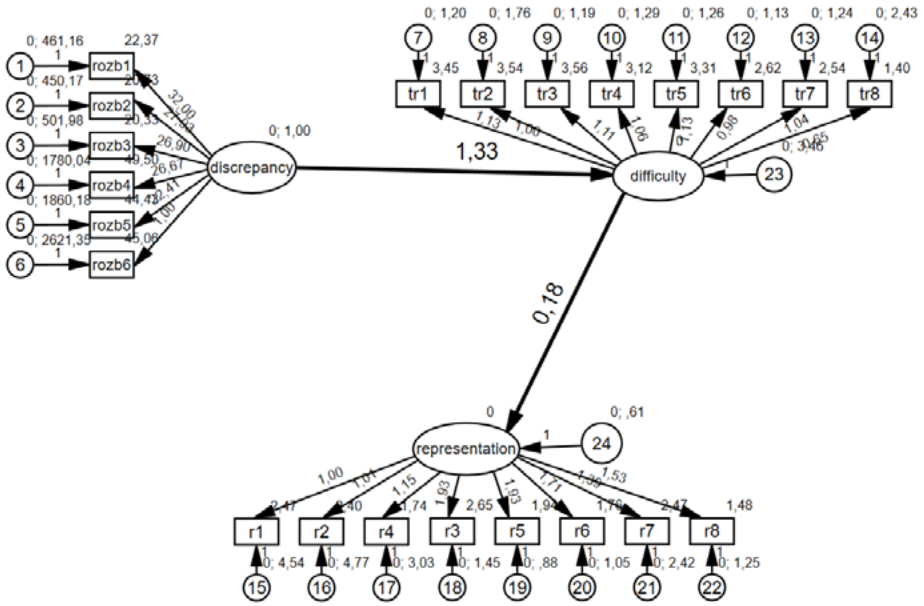


Figure 4.4. Graphical representation of the model constructed using structural equation modeling. Unstandardized results

Summarizing the analyses conducted thus far, it can be concluded that the presented model is not well-fitted to the empirical data, which undermines the validity of the theoretical model. Although all relationships in the model proved to be statistically significant and moderate in strength, the model does not meet the fit requirements.

4.3.2. Structural Model with a Hierarchical Measurement Model

Next, a structural model with a hierarchical measurement model was verified. The theoretical model was tested using structural equation modeling, retaining only those relationships that were theoretically specified. Accordingly, the path between the single-level latent structure of *Discrepancy* and the hierarchical latent structure of *Experienced Parental Difficulty* was preserved, as was the path between Experienced Parental Difficulty and the *Representation of the Child in the Parent's Mind*.

The model was estimated using the Maximum Likelihood method. The model fit results are presented in Table 4.2.

Table 4.2. Fit Statistics for the Model with a Hierarchical Measurement Model

	Fit Indices	Value	Recommended Value for Failing to Reject H_0	Statistical Significance Level
	χ^2	658.585	*ns*	$p < 0{,}001$
	df	203		
	n	319		
	χ^2*independent*	6292.969		$p< 0{,}001$
	df independent	231		
Absolute Fit Measures	Hoelther	111		$p = 0{,}05$
Relative Fit Measures – Type 1	NFI	0.891		
	RFI	0.876		
Relative Fit Measures – Type 2	IFI	0.921	>0,900	
Relative Fit Measures – Type 3	CFI	0.920	>0.900	
Fit Measures Accounting for Model Complexity	PNFI	0.783		
	PRNI			
	PCFI	0.809		
	PRATIO	0.879		
Approximation Error Measures	RMSEA	0.086	< 0,06< 0,08	90% probability of model fit

The fit measures presented in Table 4.2 indicate an adequate fit of the proposed model to the data. The value of the RMSEA statistic (close to the critical level of 0.08) and the CFI value (0.920) suggest that the presented model appropriately describes the data. This means that there is no basis for rejecting the null hypothesis, which assumes no differences between the theoretical model and the empirical model.

Figure 4.5 presents the path diagram of the estimated model, the structural part of which is represented by the following system of equations:

$$\begin{pmatrix} \text{difficulty} \\ \text{representation} \end{pmatrix} = \begin{pmatrix} \gamma_{\text{difficulty, discrepancy}} & 0 \\ 0 & \beta_{\text{representation, difficulty}} \end{pmatrix} \begin{pmatrix} \text{discrepancy} \\ \text{difficulty} \end{pmatrix} + \begin{pmatrix} \zeta_{23} \\ \zeta_{24} \end{pmatrix}$$

For the measurement model, the matrices take the form of the following system of equations:

1. For the first-order structure of the exogenous variable

$$\begin{pmatrix} \text{rozb1} \\ \text{rozb2} \\ \text{rozb3} \\ \text{rozb4} \\ \text{rozb5} \\ \text{rozb6} \end{pmatrix} = \begin{pmatrix} \lambda_{\text{rozb1, positive}} & 0 \\ \lambda_{\text{rozb2, positive}} & 0 \\ \lambda_{\text{rozb3, positive}} & 0 \\ 0 & \lambda_{\text{rozb4, negative}} \\ 0 & \lambda_{\text{rozb5, negative}} \\ 0 & \lambda_{\text{rozb6, negative}} \end{pmatrix} \begin{pmatrix} \text{positive} \\ \text{negative} \end{pmatrix} + \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \\ \sigma_5 \\ \sigma_6 \end{pmatrix}$$

2. For the second-order structure of the exogenous variable

$$\begin{pmatrix} \text{positive} \\ \text{negative} \end{pmatrix} = \begin{pmatrix} \lambda_{\text{positive, discrepancy}} & 0 \\ 0 & \lambda_{\text{negative, discrepancy}} \end{pmatrix} \begin{pmatrix} \text{discrepancy} \end{pmatrix} + \begin{pmatrix} \zeta_{25} \\ \zeta_{26} \end{pmatrix}$$

3. For the first-order of the endogenous variables

$$\begin{pmatrix} \text{tr1} \\ \text{tr2} \\ \text{tr3} \\ \text{tr4} \\ \text{tr5} \\ \text{tr6} \\ \text{tr7} \\ \text{tr8} \\ \text{r1} \\ \text{r2} \\ \text{r3} \\ \text{r4} \\ \text{r5} \\ \text{r6} \\ \text{r7} \\ \text{r8} \end{pmatrix} = \begin{pmatrix} \lambda_{\text{tr1, difficulty}} & 0 & 0 \\ \lambda_{\text{tr2, difficulty}} & 0 & 0 \\ \lambda_{\text{tr3, difficulty}} & 0 & 0 \\ \lambda_{\text{tr4, difficulty}} & 0 & 0 \\ \lambda_{\text{tr5, difficulty}} & 0 & 0 \\ \lambda_{\text{tr6, difficulty}} & 0 & 0 \\ \lambda_{\text{tr7, difficulty}} & 0 & 0 \\ \lambda_{\text{tr8, difficulty}} & 0 & 0 \\ 0 & \lambda_{\text{r1, child's tasks}} & 0 \\ 0 & \lambda_{\text{r2, child's tasks}} & 0 \\ 0 & \lambda_{\text{r4, child's tasks}} & 0 \\ 0 & 0 & \lambda_{\text{r3, own tasks}} \\ 0 & 0 & \lambda_{\text{r5, own tasks}} \\ 0 & 0 & \lambda_{\text{r6, own tasks}} \\ 0 & 0 & \lambda_{\text{r7, own tasks}} \\ 0 & 0 & \lambda_{\text{r8, own tasks}} \end{pmatrix} \begin{pmatrix} \text{difficulty} \\ \text{child's tasks} \\ \text{own tasks} \end{pmatrix} + \begin{pmatrix} \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \\ \varepsilon_{17} \\ \varepsilon_{18} \\ \varepsilon_{19} \\ \varepsilon_{20} \\ \varepsilon_{21} \\ \varepsilon_{22} \end{pmatrix}$$

4. For the second-order of the endogenous variables

$$\begin{pmatrix} \text{child's tasks} \\ \text{own tasks} \end{pmatrix} = \begin{pmatrix} \lambda_{\textit{child's tasks, representation}} & 0 \\ 0 & \lambda_{\textit{own tasks, representation}} \end{pmatrix} \begin{pmatrix} \text{representation} \end{pmatrix} + \begin{pmatrix} \zeta_{27} \\ \zeta_{28} \end{pmatrix}$$

Figure 4.5 presents the standardized results, while Figure 4.6 shows the unstandardized ones. The first hypothesis assumed that the discrepancy between the parent's parental goal and the distance between the child's current developmental level and that goal is positively related to the parental difficulty experienced. The relationship between the Discrepancy construct and Experienced Parental Difficulty proved to be statistically significant, $\gamma_{11} = 0.75$; $p < 0.0005$. Thus, the first hypothesis was confirmed—it was demonstrated that Discrepancy is significantly and strongly associated with the parental difficulty experienced. Discrepancy explains 56% of the variance in Experienced Parental Difficulty, $(0.75)^2 = 0.563$.

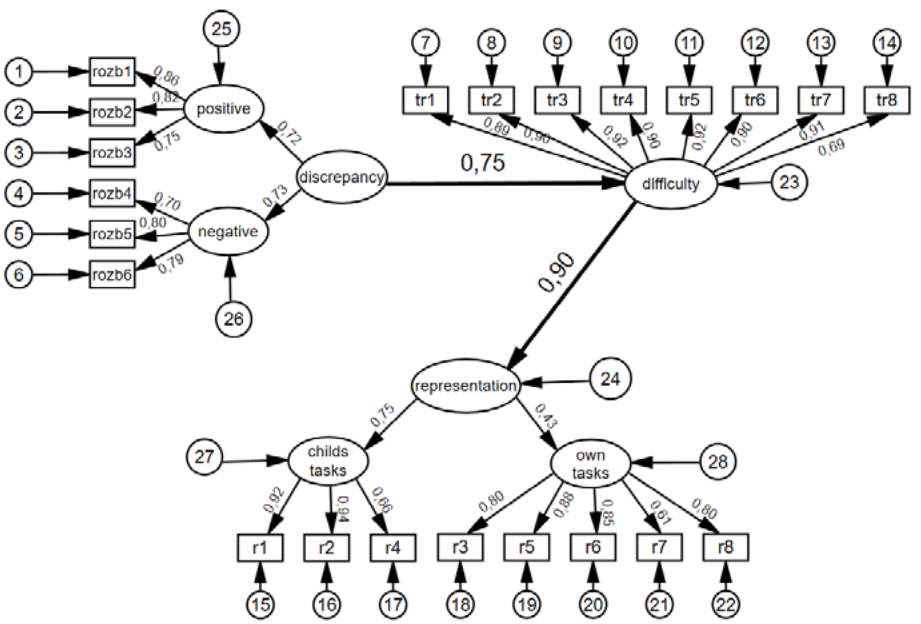


Figure 4.5. Graphical representation of the model tested using structural equations. Standardized results

The unstandardized coefficient for this relationship is 1.73, which means that each one-unit increase in Discrepancy corresponds to a 1.73-unit increase in the parental difficulty experienced (see Figure 4.6). The second hypothesis assumed a positive relationship between Experienced Parental Difficulty and the Representation of the Child in the Parent's Mind. This relationship also proved to be

statistically significant, $\beta_{21} = 0.90$; $p < 0.0005$, which confirms the validity of the second hypothesis. It was demonstrated that the relationship between these constructs is positive and strong. Experienced Parental Difficulty explains as much as 81% of the variance in the child's representation, $(0.90)^2 = 0.81$. The regression coefficient for this relationship is 0.62, meaning that a one-unit increase in difficulty results in a 0.62-unit increase in the discussed representation of the child in the parent's mind (see Figure 4.6).

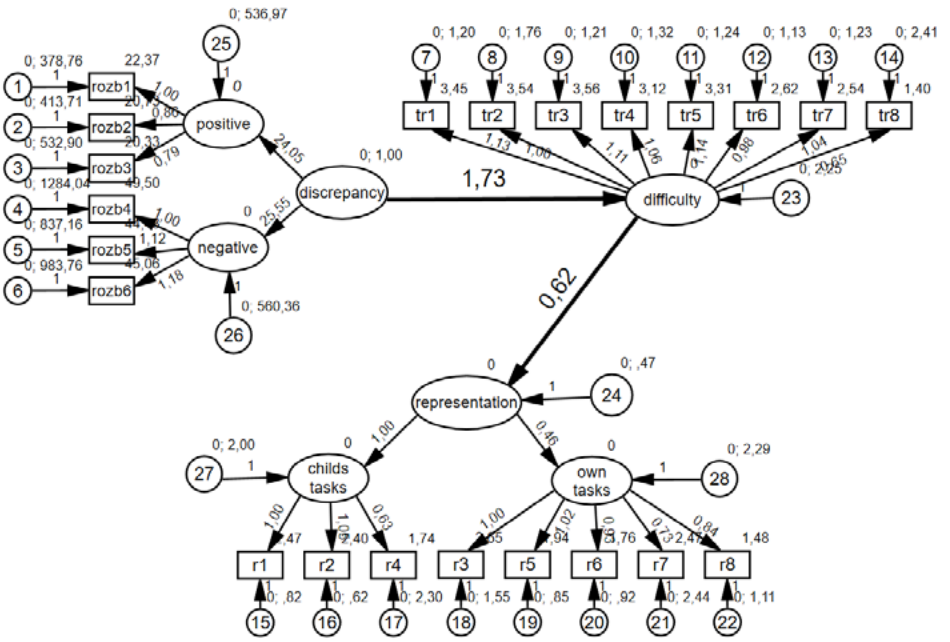


Figure 4.6. Graphical representation of the model tested using structural equations. Unstandardized results

Summarizing the analyses conducted thus far, it should be concluded that the structural model with a hierarchical measurement model is well-fitted to the empirical data, which confirms the validity of the proposed theoretical model. All relationships in the model proved to be significant and strong. It is worth noting that the values of both standardized and unstandardized coefficients increased substantially in comparison to the model with a single-level measurement model.

For comparison, the standardized relationship between Discrepancy and Experienced Parental Difficulty in the model with a hierarchical model was $\gamma_{11} = 0.75$; $p < 0.0005$, whereas in the model with a single-level model this value was $\gamma_{11} = 0.58$; $p < 0.0005$. The standardized relationship between Experienced Parental Difficulty and the Representation of the Child in the Parent's Mind in the hierarchical model was $\beta_{21} = 0.90$; $p < 0.0005$, whereas in the single-level model it was $\beta_{21} = 0.46$; $p < 0.0005$.

The notable increase in these coefficients results from the fact that the model with a hierarchical measurement model more accurately reflects the structure of the latent variables, which directly translates into higher values of lambdas (factor loadings). The lower lambda values in the single-level model suggested that the latent variables explained less of the variance in observable variables, which led to greater error variance and reduced coefficient values. This had a significant impact on the model fit statistics, ultimately resulting in the rejection of the single-level model.

It should be emphasized that adherence to modeling assumptions is crucial in model construction. Proper selection of indicators for latent structures is essential for model validity, and it is equally important to correctly specify the paths between latent variables.

CHAPTER 5

Analysis and Interpretation of Structural Model Results

5.1. The Idea of Mathematical Modeling

This chapter devotes special attention to the idea of mathematical modeling, which constitutes the foundation of structural equation models (SEM). Mathematical modeling in SEM is based on the concept of comparing two key matrices: the *saturated matrix* and the *default matrix*. Understanding what these matrices represent and how they are compared is essential for accurate interpretation of the results (Hair et al., 2006).

The saturated matrix represents a full model in which all possible relationships between variables are accounted for. In other words, this model contains the maximum number of parameters, which results in a perfect fit to the data. In mathematical modeling, the saturated matrix serves as a reference point, enabling the evaluation of the extent to which simplified models reflect the relationships among variables found in the data.

The default matrix, also known as the received matrix, differs from the saturated matrix in that not all parameters are estimated. This means that not all possible relationships between variables are assumed. Such a simplification of the model increases the number of degrees of freedom, which, on the one hand, allows for a more economical use of parameters but, on the other hand, carries the risk of oversimplifying the model.

A key element of SEM result analysis is the comparison between the saturated and default matrices. Using the Chi-square statistic, one assesses the extent to which the default matrix differs from the saturated matrix. In cases where the difference

between these two matrices is too large, one may conclude that the adopted model simplification was inappropriate, which in turn may lead to incorrect conclusions regarding the investigated relationships (Szymańska, 2016b).

Below is an example of two square matrices, each of size four-by-four, illustrating the difference between a saturated matrix and a default matrix.

Saturated Matrix

1.0	**0.8**	0.5	**0.7**
0.8	**1.0**	0.4	0.6
0.5	0.4	**1.0**	**0.9**
0.7	0.6	**0.9**	**1.0**

Default Matrix

1.0	**0.8**	0.0	**0.7**
0.8	**1.0**	0.0	0.0
0.0	0.0	**1.0**	**0.9**
0.7	0.0	**0.9**	**1.0**

In the example above, the **saturated matrix** contains a complete set of parameters, meaning that all possible relationships between variables have been accounted for. Each cell in the matrix represents the strength of the relationship between two variables. As shown, there are no empty cells or zero values, which indicates that all relationships have been included.

In contrast, in the **default matrix**, some relationships have been deemed insignificant or were omitted from the model, which is reflected by the presence of zero values in the respective cells. This representation reflects a situation in which certain relationships were not included in the model as a result of simplifying the model by excluding selected parameters, thereby increasing the degrees of freedom.

In practice, such model simplification—leading to the construction of the default matrix—may aid in achieving a more economical model, but it can also result in the loss of important information about the relationships between variables if the excluded parameters correspond to actual associations present in the data. A comparative analysis of these two matrices allows for an evaluation of whether the simplifications introduced in the model are justified.

As can be observed, all relationships in the saturated matrix that are smaller than 0.7 were not included in the default matrix. In the default matrix, these values were replaced with zeros, indicating that these weaker relationships were considered insignificant enough to be excluded from the model. The relationships that appear in both the saturated and the default matrices have been bolded in the matrices.

Accordingly, the Chi-square statistic will compare these two matrices by assessing how the omission of these weaker relationships (with values of 0.5, 0.4, 0.6)

affected the model fit. The Chi-square analysis evaluates the extent to which the default matrix differs from the saturated matrix. If the differences are too large—that is, if the omitted relationships are actually meaningful in the data—then the Chi-square statistic may indicate that the model is misfitting. In that case, such a model would have to be regarded as inadequately fitted to the data.

It is also important to understand that in the case of simple models with a small number of parameters, even considerable differences in a few places may not lead to a high Chi-square value, which means the model may still be accepted. However, in the case of more complex models, where the number of parameters is significantly larger, even small differences—such as values on the order of 0.2 or 0.3—may accumulate into hundreds or thousands of minor discrepancies. In such situations, the Chi-square value can increase dramatically, leading to the rejection of the model, even if the model appears to be theoretically sound.

This explains why simple models are often not rejected by the Chi-square test, whereas more complex models—even if well-constructed—may be rejected due to a larger number of estimated parameters and a lower number of degrees of freedom. Therefore, it is crucial in mathematical modeling within SEM to pay attention to the number of parameters and the degree of model simplification, bearing in mind the potential risk of model misfit.

For this reason, to better assess model fit, other measures are used that do not rely solely on the Chi-square statistic. These additional indices take into account not only the model fit but also its complexity. Examples of such measures include RMSEA (Root Mean Square Error of Approximation) and CFI (Comparative Fit Index), which introduce an additional level of control over model complexity by incorporating the number of degrees of freedom and comparing the tested model to an independent model.

RMSEA evaluates how well the model fits the data, taking into account both statistical fit and model complexity. The more parameters are estimated, the fewer the degrees of freedom and the higher the risk of overfitting the model to the data. RMSEA attempts to correct this tendency by penalizing models with a low number of degrees of freedom that incorporate too many relationships between variables.

RMSEA is a measure that not only assesses how well the model fits the data but also considers its complexity. The RMSEA value accounts for the number of degrees of freedom, that is, the degree of model simplification—as demonstrated in Equation 4.1. Models with a higher number of degrees of freedom are simpler because they assume fewer relationships between variables. As a result, RMSEA penalizes overly complex models, i.e., those that include many parameters and have a low number of degrees of freedom.

The penalty in RMSEA consists in the fact that models containing a large number of estimated parameters—thus having few degrees of freedom—may receive a higher RMSEA value if their fit to the data is insufficient. A high RMSEA value signals that the model's complexity does not translate into an accurate representation of the relationships between variables. Overfitting means that the model may

be representing random fluctuations in the data too precisely, rather than capturing genuine relationships. In practice, such a model may be less useful when applied to the analysis of new data.

Therefore, RMSEA is a valuable tool, as it helps balance the accuracy of model fit with its complexity. Models that achieve low RMSEA values are generally more balanced—they fit the data well while avoiding excessive complexity. This enables a more accurate assessment of a model, in which complexity does not overshadow actual data fit.

In summary, RMSEA penalizes models for excessive complexity, which means that simpler models—with a greater number of degrees of freedom, i.e., estimating fewer parameters—are generally evaluated more favorably by this measure. As a result, the researcher can better assess whether a given model is overly complicated and whether it appropriately reflects the actual relationships between variables.

In turn, the CFI (Comparative Fit Index) compares the fit of the tested model (*default model*) with that of an *independence model*, which assumes that all variables are entirely independent from one another (Hair et al., 2006). This index is based on the analysis of relative improvement in fit—it assesses how much better the default model represents the data compared to a model that assumes no associations between variables. A high CFI value indicates that the *default model* fits the data well, even when accounting for the number of estimated parameters, i.e., its complexity.

These additional indices allow for a more balanced evaluation of a model, as they take into account both fit and complexity. Thanks to them, it is possible to avoid situations in which complex models—although they represent the data well—are rejected solely based on the Chi-square statistic. The use of such measures as RMSEA and CFI is therefore a key component of model analysis in SEM, allowing for more accurate conclusions about how well the model reflects reality.

The more complex the model, the higher the risk that its fit will be overly tailored to a specific dataset, limiting its capacity to generalize results and to predict in other contexts. Such a model becomes specific to the analyzed dataset, which reduces its predictive value and its usefulness in broader research.

One of the key assumptions in modeling is the treatment of measurement errors as random and independent (Hair et al., 2006). Introducing correlations between measurement errors is highly problematic, as it suggests the existence of an uncontrolled factor not accounted for in the model. This implies that the errors are no longer random, which points to serious shortcomings in the construction of the measurement model.

Correlating measurement errors leads to an artificial fit of the model to the data, which distorts the actual relationships between variables (Hair et al., 2006). Such manipulation of the model is unacceptable, as it falsifies the representation of reality and leads to erroneous conclusions. For this reason, avoiding the correlation of measurement errors should be treated as one of the fundamental principles when constructing models aimed at reliably representing the phenomena under investigation.

In practice, this means that researchers must exercise particular caution to ensure that their models do not introduce artificial correlations between measurement errors, as this could indicate a flawed understanding of the phenomena being analyzed. A model that fits the data well solely by correlating measurement errors is a poorly constructed model and does not deserve to be regarded as credible.

In structural equation modeling (SEM), there is a specific situation in which a model, although statistically well-fitted to the data and accounting for freed degrees of freedom, does not provide any meaningful conclusions about the investigated phenomena. This occurs when the relationships between variables in the model are very weak or nearly independent.

If all correlations in the saturated matrix are low, any simplification of this model by removing certain parameters from the saturated matrix does not lead to significant changes in the default matrix. As a result, the default matrix remains very similar to the saturated one, which may lead to high values of fit indices such as the CFI (*Comparative Fit Index*). However, a high CFI value in such a model is misleading because it suggests good fit, even though the model does not provide any substantive conclusions.

Models characterized by weak relationships between variables tend to achieve high fit indices precisely because the difference between the saturated and default matrices is minimal. Meanwhile, models in which variables are strongly related may yield poorer fit indices, even though they provide significant information about actual relationships in the data.

Therefore, one must be cautious when interpreting fit indices, especially in cases where models exhibit weak relationships between variables. A high CFI value is not always an indicator of a good model; it may merely reflect a low difference between the saturated and default matrices, resulting from weak associations in the data. Consequently, models with strong relationships, although they may have lower fit indices, may better reflect reality and yield more valuable conclusions.

These and similar issues become more apparent once the idea of modeling is clearly understood—which is precisely why this chapter begins with that concept. The following sections will explore other aspects of modeling.

5.2. Modeling Structural Relationships

Causal relationships constitute the foundation of structural models, often referred to as substantive models. The structural theory underlying these models not only describes phenomena but also the relationships between various elements, which directly relates to causal processes. For this reason, structural models are fundamentally based on processual phenomena—namely, cause-and-effect relationships (Konarski, 2009).

However, a significant methodological issue arises. Structural equation modeling (SEM) is not an experimental method, but rather one based on correlational techniques. It has long been recognized that correlational methods cannot definitively

establish causal relationships. This raises the question: can SEM, despite being based on theories describing causal phenomena, be used to verify such relationships?

The answer is: not directly. A structural SEM model does not confirm causal phenomena—it merely *makes them more probable*. In other words, when a model that describes a causal phenomenon proves to be statistically valid, this increases the likelihood that the theory describing the phenomenon is accurate. However, this does not amount to a confirmation of the theory in the causal sense, but rather a probabilistic support for it (Szymańska & Aranowska, 2016).

What, then, allows SEM to lend plausibility to causal phenomena? The key lies in theory. It is the theory preceding the construction of the structural model that serves as the foundation for that model. Because SEM is based on a previously formulated theory, we may infer that, if the model is statistically valid, it increases the likelihood that the theory describing these phenomena is also valid. The theory confers status and value upon the model, making SEM a tool that can support causal inference, though without providing definitive confirmation (Szymańska, 2016b; Szymańska & Aranowska, 2016).

The role of theory in methodological research cannot be overstated. It is theory that enables us to speak at all about the plausibility of causal relationships using structural equation modeling. Theory gives meaning and direction to our models, making them valuable tools in the research process.

In the context of structural models that are not based on a previously formulated theory but are instead derived exploratorily from data, an important question arises regarding their value and role in research. These models—often referred to as alternative models or models "dug out" from the data (as Jan Gajda describes them)—hold a different status than models grounded in robust theoretical foundations (Gajda, 1992).

Alternative models that lack an underlying theory cannot provide support for causal relationships. Their construction did not stem from a theory describing the causes and effects of the phenomena under study. Instead, they were derived from data analysis, which means they reflect only the structures and relationships present within the analysed sample. As such, their role is purely exploratory—they describe what is contained in the data but do not provide strong grounds for causal inference.

Due to the absence of theory, these models hold a lower rank in the hierarchy of scientific research. They do not serve a confirmatory role but may act as starting points for further investigation. In practice, this means that models "dug out" from data require additional confirmation on independent research samples. Their value lies in identifying potential relationships that may be of interest, but they are not capable of independently confirming theories or supporting causal claims (Gajda, 1992).

Therefore, such models should be treated as exploratory tools that highlight areas requiring further study. Their value lies in the generation of new hypotheses, which may then be tested in research based on stronger theoretical foundations.

Only such verification on new samples can lead to the construction of theories and models that possess greater value in terms of supporting causal inference.

Among the alternative models, there is also a group of models which, although not "extracted" from the data, are just as significant as the primary model. These alternative models likewise describe structural relationships, but their foundation lies in theory rather than data exploration alone. It often happens that a theory predicts several possible paths through which a given process might unfold, leading to the formulation of more than one structural model (Hair et al., 2006; Jonkisz, 1998; Szymańska, 2016b).

In structural modeling, it is rarely possible to predict all possible combinations of relationships among variables; therefore, multiple models may emerge. This may include a primary model and one or more alternative models that are also theoretically derived. The academic literature not only allows for, but in fact encourages, the construction of such alternative models, as they enable a more comprehensive understanding of the phenomenon under investigation (Hair et al., 2006).

In cases where alternative models are derived from theory, they carry the same scientific weight as the primary model. The condition is that both the primary and the alternative model are rooted in a sound theoretical framework. In such cases, both models are treated with equal seriousness, as both may support causal inferences—provided the theory on which they are based is well-established.

In summary, not all alternative models are created equal. Alternative models grounded in theory hold the same scientific value as the primary model, because theory provides them with the necessary legitimacy to investigate causal phenomena (Szymańska & Aranowska, 2016). For this reason, the researcher should treat both the primary and the alternative theoretical models with equal consideration, as each may offer a different but equally important perspective on the phenomenon being studied.

The issue of model simplicity plays a key role in evaluating the scientific merit of models and is closely related to the principle known as *Occam's Razor* (Hair et al., 2006; Szymańska, 2016b). This principle, also referred to as the principle of parsimony or economy, originates from the medieval philosopher and theologian William of Ockham (c. 1287–1347). Although Ockham himself never used the term "Occam's Razor", his ideas laid the foundation for the principle, which suggests that among competing explanations of a phenomenon, the one that requires the fewest assumptions should be preferred. This principle not only simplifies explanations but also enhances their practicality and testability, which is essential in scientific methodology (wikipedia contributors, 2024).

Since the Middle Ages, this principle has been used in science as a tool to facilitate the evaluation of competing theories. It is widely accepted as a directive according to which the better model is the one that explains phenomena equally well but is simpler—that is, based on fewer assumptions. In science, Occam's Razor leads to a preference for theories that are easier to test and less prone to error (Szymańska, 2016b).

In the context of structural modeling, this principle has direct application. When a researcher constructs a structural model and then compares it with alternative models, they should aim to select the simplest model among those that explain the phenomenon equally well. This principle underpins the directive to construct alternative models in order to search for the best, most parsimonious explanation. However, this does not imply a preference for simplified models at the expense of theoretical accuracy—parsimony pertains exclusively to structural economy while maintaining substantive adequacy.

In practice, if we have two models that explain a given phenomenon equally well, the simpler one should be preferred. Such a model not only better meets methodological criteria but is also generally more useful in practice, as simplicity facilitates its verification and application in other contexts (Szymańska, 2016b).

It is worth noting that although Occam's Razor is extremely helpful, it is not without limitations. Excessive simplification of a model may lead to underfitting, which is why it is important to maintain a balance between simplicity and explanatory adequacy. Nevertheless, as a general methodological rule, Occam's Razor remains a cornerstone in evaluating the scientific value of models, especially in the field of modeling structural relationships (Szymańska, 2016b).

Emphasising the principle of simplicity and parsimony is crucial in the context of mathematical modeling, because every model—regardless of how well it fits—is always a simplification of reality. It is important to remember that our pursuit of the simplest possible models is justified not only by methodological principles but also by the fundamental nature of modeling itself.

Every mathematical model, including structural models, is incapable of fully reflecting the complexity of the real world. This results from the fact that reality is exceedingly complex and multidimensional, and no theory, however advanced, can capture all its aspects. Every theory, as well as the model derived from it, necessarily selects only those aspects of reality deemed most relevant for describing a given phenomenon (Jonkisz, 1998).

For this reason, modeling involves selection and simplification—identifying those elements that are essential for understanding and predicting the course of certain processes. However, even at the stage of constructing a model, the researcher must remain aware that every model is a simplification of reality and cannot reflect the full complexity of the phenomena it seeks to describe. From this perspective, a model does not precisely replicate reality, but rather suggests which processes occur most frequently and which elements play the most significant roles within them.

Therefore, simplicity in modeling is not only a goal but a necessity. Models must be simplified in order to be useful and comprehensible, but this simplicity will always involve a kind of compromise—a relinquishment of a full description of reality in favour of capturing its most essential relationships. As a result, even the best-constructed model does not offer absolute truth about reality, but rather the most probable representation of how certain processes may unfold, taking into account the most important—not all possible—elements (Hair et al., 2006).

In light of the above considerations, it is worth emphasising that the academic literature frequently points out that at the foundation of every mathematical model lies an already false null hypothesis. In his work, for example, Konarski (2009) argues that this is unavoidable due to the very nature of modeling.

Why is the null hypothesis false? Because a model, as a simplification of reality, does not fully reflect the truth about that reality. As mentioned earlier, every model must simplify complex phenomena by focusing on key elements while omitting others. Therefore, the null hypothesis on which the model is based is necessarily simplified to such a degree that it fails to capture the full complexity of the studied phenomena.

Konarski (2009) notes that the falsity of the null hypothesis is not a flaw of the model, but rather a natural outcome of the modeling process. Mathematical models are intended to provide useful and practical tools for analysing reality, but their effectiveness does not lie in perfectly mirroring the world, but rather in their ability to capture key relationships in a comprehensible and applicable manner. The null hypothesis in a model is thus false in the sense that it is a simplification, but at the same time, it is an essential element that enables the model to be used in scientific research.

As a result, researchers must remain aware that every null hypothesis and every model based on it are merely approximations of reality, serving to test and support the plausibility of theories rather than to confirm them absolutely.

The structural equation model (SEM) differs fundamentally from traditional regression or correlation models because it is not limited to analysing simple linear relationships between two variables. SEM is based on the analysis of covariance matrices, which allows for the modeling of complex, directional relationships among multiple variables simultaneously. This enables the investigation of both observable and latent variables, as well as the inclusion of measurement errors.

One of the key differences between SEM and linear regression is that in SEM, reversing the direction of the relationship between variables may lead to different results—something that does not occur in simple regression models. In a regression model, the relationship between variables A and Z is symmetric, meaning that the direction of analysis (i.e., whether we examine the influence of A on Z or Z on A) does not change the results. In contrast, in SEM, such relationships are part of a more complex model structure, meaning that altering the direction of one relationship may affect other dependencies within the model, leading to different conclusions.

This distinction is illustrated by a model presented in the book *Upbringing Mistake: Toward the Verification of Antonina Gurycka's Theory* by Szymańska and Aranowska (2016). In the original model from the book, the relationship between the variable "child representation" and "aggressive directiveness applied by the parent" was unidirectional. In a new analysis, an additional arrow was introduced, reversing the direction of the relationship, which allowed for the mutual influence of these variables to be considered. A diagram presenting the results of this modified version of the model is shown in Figure 5.1.

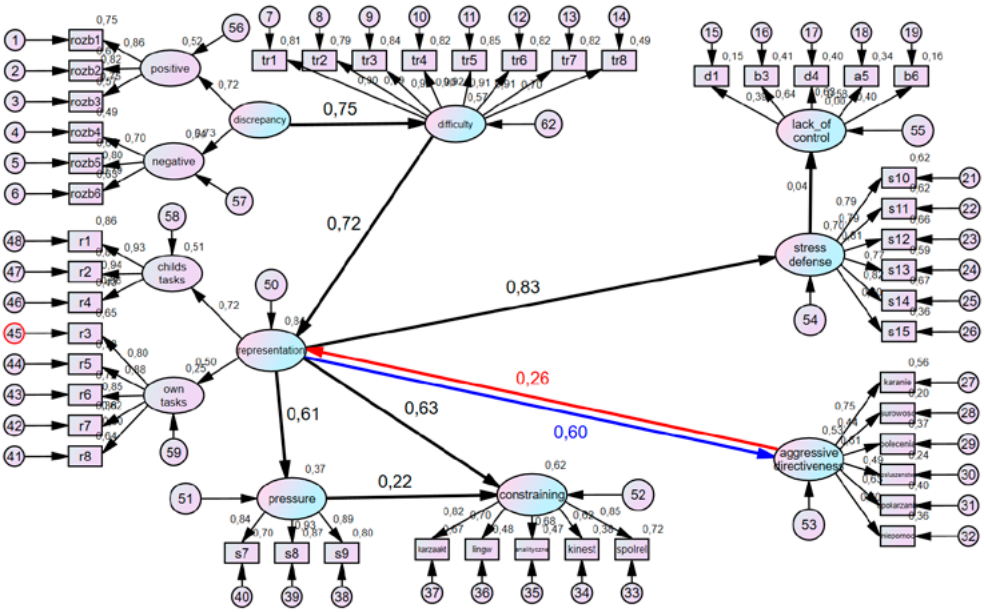


Figure 5.1. Non-recursive model with a bidirectional relationship between the variables "child representation in the parent's mind" and "aggressive directiveness"

As the analysis shows, when the effect of child representation on aggressive directiveness is examined, the value of the beta coefficient is 0.6. However, when the direction of the relationship is reversed—i.e., when the effect of aggressive directiveness on child representation is analysed—the beta coefficient decreases to 0.26. This example clearly demonstrates that the relationship between two variables may vary depending on the direction of the analysis, which is a unique feature of SEM and highlights its complexity and flexibility in modeling causal phenomena.

In summary, the structural equation model not only enables the examination of complex relationships among variables, but also allows for the testing of directional dependencies, which would be treated symmetrically in traditional regression models. As a result, SEM offers a more realistic analytical tool that accounts for the complexity of real-world phenomena, which is particularly important in research on causal relationships.

To conclude, although the introduction of artificial intelligence and its algorithms represents a significant advancement in data analysis, structural equation models remain irreplaceable. Their role in the verification of theoretical models is unquestionable, and their application in scientific research remains uncontested. Therefore, even in the face of emerging technologies, SEM retains its central position in research methodology, offering a tool that, despite certain limitations, is indispensable for the analysis of complex phenomena.

5.3. Limitations of Models Verified by Means of Structural Equation Modeling

This chapter discusses the limitations of models verified using structural equation modeling (SEM) and the necessity of introducing new methods of analysis, such as artificial intelligence (AI) and AI algorithms, particularly in the context of processual phenomena. At the outset, it is essential to emphasise the tremendous significance of SEM, which remains unquestionable. Structural equation models play a key role in the verification of theoretical models and, despite the emergence of new technologies, do not lose their relevance.

SEM serves as a foundational research tool, enabling scholars to test complex theories by modeling relationships among variables, both observable and latent. Due to their capacity to integrate multiple variables and account for measurement errors, these models are indispensable in the study of social, psychological, and educational phenomena. It is precisely this complexity and precision that make SEM impossible to ignore—even in the era of rapidly advancing artificial intelligence technologies.

However, like any tool, SEM has its limitations. These models are highly dependent on theoretical assumptions and the quality of the input data (Aranowska, 2005). The assumptions adopted may at times fail to fully reflect reality, leading to simplifications and potentially erroneous conclusions. Moreover, SEM is a deductive tool, meaning its effectiveness relies on the accuracy of the theory being tested. In cases where the theory is not sufficiently developed, the results of SEM analysis may be limited (Hair et al., 2006).

In this context, the role of artificial intelligence and AI algorithms is growing, offering new possibilities for data analysis, particularly with regard to processual phenomena (Nisbet et al., 2009). With its ability to process vast amounts of data and to identify patterns that may not be evident in traditional models, AI constitutes a valuable complement to SEM. However, at least for now, AI does not appear capable of replacing SEM but rather serves a supporting role in the modeling of psychological phenomena.

One of the significant limitations of SEM is that, although it reveals general relationships within a population, it is not always capable of capturing differences in these relationships across subgroups within that population. For example, an SEM model may indicate a strong relationship between a parental mistake and a discrepancy in parental values as well as difficulty in the parent–child relationship. However, this relationship may pertain only to a specific segment of the population—e.g., 15–20%—while for other groups it may be considerably weaker or even non-existent (Szymańska, 2019).

In such cases, SEM provides only a general overview, without showing how different groups may vary in terms of the intensity and nature of the phenomena under study. Therefore, it is advisable to complement SEM analysis with additional

methods that enable a more detailed understanding of population structure. For instance, the use of cluster analysis, derived through artificial intelligence algorithms, may help identify and interpret specific subgroups for which relationships differ (Szymańska, 2019).

This approach not only addresses the limitations of SEM but also allows for a more precise understanding of the phenomena studied. Cluster analysis may reveal that relationships appearing strong at the population level, in reality, concern only a small segment, which is crucial for accurate interpretation of results. As a result, we obtain a more complete picture and can provide more precise answers to research questions (Szymańska, 2018, 2019).

In response to the limitations stemming from the fact that structural equation models (SEM) show only general relationships within a population and do not account for differences across subgroups, more advanced approaches have been developed within the SEM framework, such as multilevel structural equation models (MSEM). These models were introduced to account for the fact that relationships at the basic level may be stronger in some subgroups and weaker in others.

As noted by Heck and Thomas (2009), hierarchical structural equation models allow for the modeling of relationships between variables at different levels of analysis, taking into account population heterogeneity. This enables a more precise understanding of how particular phenomena operate across various contexts, considering differences in subgroups that may remain hidden when data are analysed only at the general level (Heck & Thomas, 2009).

MSEM facilitates the analysis of relationships between variables at different levels—for example, individual and group levels—which is particularly useful in studies where there is a need to understand how contextual variables influence relationships at the individual level. As a result, multilevel models provide a more complex and at the same time more realistic picture of the phenomena under investigation.

This approach allows for better alignment of models with reality, in which not all relationships are equally strong across the entire population. In this way, hierarchical SEM becomes a powerful tool that enables deeper analysis and better understanding of behavioural dynamics across different groups and contexts.

SEM is an extremely valuable tool for examining relationships among variables and verifying theoretical structures. However, it is important to note that SEM is not a predictive technique per se. SEM models primarily focus on confirming relationships among variables and assessing whether the theoretical structure fits the data well. It is assumed that models with strong relationships between variables and good fit should also demonstrate high predictive power, but in practice, this is not always the case.

Prediction in SEM is rather a by-product of a well-fitting model than its primary aim. This means that while SEM models may suggest how variables are likely to behave in the future, they are not designed for forecasting phenomena in the same way as specialised predictive models. Therefore, assessing whether a theory and its

structural model possess predictive capabilities should be supplemented with other methods, such as predictive models based on artificial intelligence algorithms—for example, artificial neural networks.

Szymańska highlights that combining SEM with artificial neural networks (ANNs) may enable a better understanding of the extent to which structural models possess predictive properties. While SEM focuses primarily on confirming relationships among variables and evaluating the fit of theoretical models, integration with ANNs allows for the assessment of how well these models can predict future phenomena. The application of ANNs enables data analysis in a way that complements SEM—particularly in the context of prediction—allowing researchers not only to verify theories but also to evaluate their predictive capabilities in practice (Szymańska, 2018, 2019).

In summary, one of the limitations of structural equation modeling is its lack of direct predictive capacity. Although SEM allows for the assessment of the congruence between theory and data, its predictive outcomes rely on the assumption that well-fitting models will also be good predictors of phenomena—an assumption that is not always guaranteed (Szymańska, 2018). Therefore, it is recommended to supplement SEM with predictive methods in order to fully evaluate the extent to which a given model and theory may be applied to the prediction of real-world phenomena.

Structural equation models (SEM) offer a tool for exploring complex relationships among variables; however, constructing exploratory models in SEM involves significant challenges. One of the main difficulties faced by researchers is the need to select, from a multitude of interrelationships, those that are most important and those that may be omitted (Hair et al., 2006). This process is complex and requires substantial competence and experience in evaluating which relationships should be estimated and which should not.

In such cases, algorithms—though not necessarily originating from the field of artificial intelligence—may prove highly useful in supporting the model-building process. An example of such a tool is the algorithm of grade correspondence analysis, which specialises in optimising matrix structure by maximising the correlation between rows and columns, corresponding to the variables arranged within those structures (Jarochowska, 2005a). This algorithm assists the researcher in identifying an optimal solution for the model, minimising the need for manual selection of variables and relationships—one of the most time-consuming and complex tasks in data exploration.

This book presents an original solution that illustrates how this algorithm can support the construction of exploratory models in SEM. This is particularly important when the researcher is dealing with highly complex datasets, where manually selecting the key relationships would be practically impossible.

In summary, although structural equation models are a powerful tool in data analysis, the construction of exploratory models remains one of their primary limitations. Algorithms such as the grade algorithm can significantly improve this process by helping researchers more efficiently identify key relationships in exploratory models.

The final limitation discussed here in the context of structural equation models (SEM) pertains to their statistical constraints. SEM requires the fulfilment of a number of assumptions that must be met at the data level in order for the method to be applied. These limitations are related both to the properties of the data and to the construction of the model itself (Aranowska, 1996, 2005).

First, SEM assumes that the data are linear in nature and that the relationships between variables are also linear. This means that SEM is not suitable for analysing nonlinear relationships, which presents a significant limitation in cases where the actual associations between variables deviate from linearity (Hair et al., 2006).

Second, there are requirements concerning the measurement scale of variables. SEM performs best with quantitative data, where variables are measured on interval or ratio scales. Qualitative data, such as nominal or ordinal categories, are difficult to analyse using SEM (as they require appropriate estimators), which limits the applicability of this technique in such datasets (Rosseel, 2012).

Another key assumption is homogeneity of variance (homoskedasticity), which posits that error variances are equal across all observations. Violation of this assumption may lead to inaccurate results and interpretations in SEM (Hair et al., 2006).

SEM also requires that data follow multivariate normal distributions. When data distributions significantly deviate from normality, SEM results may be invalid or difficult to interpret (Aranowska, 1996). The same applies to distributions that violate assumptions of homogeneity of variance or *lack of autocorrelation.*

Due to these requirements, SEM is not suitable for analysing data that do not meet these assumptions. These models are therefore applicable only in cases of linear relationships that also satisfy strictly defined statistical conditions. These limitations mean that, despite its strength as an analytical tool, SEM has clear boundaries in its applicability and often requires supplementation with other methods, such as artificial intelligence algorithms.

This book will demonstrate how AI algorithms can be used to address the limitations of SEM, offering alternative approaches that may assist in analysing more complex and nonlinear relationships that do not meet traditional SEM assumptions. In this way, it becomes possible to achieve a fuller and more comprehensive picture of the phenomena under investigation.

PART II

Inductive Algorithms Creating Rules for Building Decision Trees

CHAPTER 6

Fundamentals of Modeling with Decision Trees

The algorithms discussed in this chapter belong to the group of inductive algorithms. Their main purpose is to discover rules and create generalizations based on the analyzed data, which can then be used in subsequent stages of data analysis (Dramiński, 2007). The results produced by these algorithms are presented in the form of a decision tree—a structure that enables the division of a dataset into homogeneous classes, such as groups of individuals with similar characteristics.

The construction of the tree is based on creating the simplest possible structure, characterized by a minimal number of nodes. This approach allows for the generation of clear and easily interpretable decision rules. An overly complex tree structure may hinder interpretation and lead to overfitting; therefore, it is essential to maintain a balance between accuracy and model transparency.

The structure of a decision tree consists of the following elements:

- branches (edges), which connect elements of the tree,
- nodes (vertices with at least one outgoing edge),
- leaves (vertices with no outgoing edges).

Figure 6.1 presents an example illustrating these elements.

Decision trees are not only tools for formulating rules but also enable the evaluation of predictive performance on other datasets. Due to their versatility and interpretability, decision trees—like artificial neural networks—are widely used in the construction of expert systems (Michalik, 2006a; Rutkowski, 2006; Tadeusiewicz, 2012).

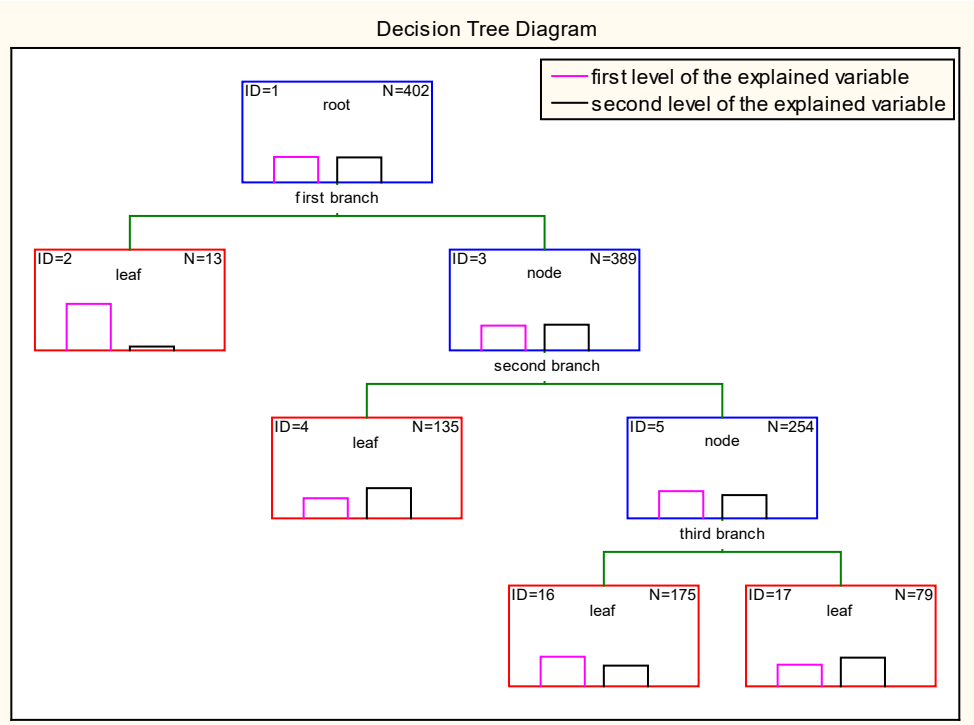


Figure 6.1. Diagram of a Decision Tree

Predictive methods based on decision trees include two main types: **classification trees** and **regression trees**. The criteria for data splitting in these trees are based on maximizing group homogeneity (e.g., using the Gini index or entropy).

The algorithm for constructing decision trees follows these steps:

1. Specification of accuracy criteria – defining the required performance level for the model's predictions.
2. Searching for split points – identifying locations that allow the dataset to be most effectively divided into homogeneous groups.
3. Determining the stopping point for category splitting – the C&RT algorithm initially builds a maximally complex tree, which may be overly fitted to the data. Overfitting means that the results cannot be easily generalized to the population, as they depend heavily on the specific features of the studied sample.
4. Deciding on the optimal tree size – to avoid overfitting, tree pruning is applied, using cross-validation. This process involves evaluating the error for each tree branch and checking whether simplifying the structure (treating the root of a branch as a leaf) reduces the model's overall error.

Decision trees can be built for both continuous dependent variables (in which case they are called *regression trees*) and categorical dependent variables (in which

case they are *classification trees*). In classification algorithms, two types of predictions are distinguished:

1. Classification prediction – assigning an object to one of the predefined groups.
2. Distributional prediction – determining the probability of the dependent variable's value occurring as a result of the independent variable's influence.

Error and Uncertainty Measures in Classification Trees

In decision tree models, various indicators are used to evaluate classification performance. These can be divided into measures of classification errors and measures of tree structure uncertainty.

In classification trees, two types of error are distinguished:

1. Classification Errors:

- **Misclassification rate** – indicates the percentage of objects that were assigned to the wrong group. This is the basic accuracy measure of the model, treating all errors as equally important.
- **Average loss rate** – accounts for different costs of errors. For example, in aircraft diagnostics, misclassifying a malfunctioning aircraft as operational is far more costly than the reverse. An example decision cost table is presented in **Table 6.1**. The loss resulting from misclassifying a malfunctioning aircraft as operational is 1000 times greater than the loss from misclassifying an operational aircraft as malfunctioning. This table illustrates the asymmetry of decision costs, which is particularly important in high-responsibility decision systems, such as aviation.

Table 6.1. Frequency Table for the Cost of Wrong Decisions

Decision / Actual Status	Malfunctioning	Operational
Malfunctioning	0	1
Operational	1000	0

2. Uncertainty Measure: Entropy

Data split entropy is an index that describes the level of uncertainty in classifying objects after dividing the dataset. Lower entropy indicates more homogeneous groups and greater assignment certainty. Although entropy does not directly measure classification errors, its value reflects how "clean" and organized a split is.

Entropy is closely tied to the concept of information—the lower the entropy, the higher the *information gain* resulting from a particular split. Therefore, tree-building algorithms favor those attributes that lead to splits with maximal entropy reduction.

If object classification is correct, the groups are clearly distinct and internally homogeneous. Such a division supports low error rates and a high level of model fit. On the other hand, when the data contain many ambiguous or difficult-to-classify instances, error indices rise and fit statistics decline.

Example

Figure 6.2 presents a sample decision tree computed using the C&RT (Classification and Regression Tree) algorithm.

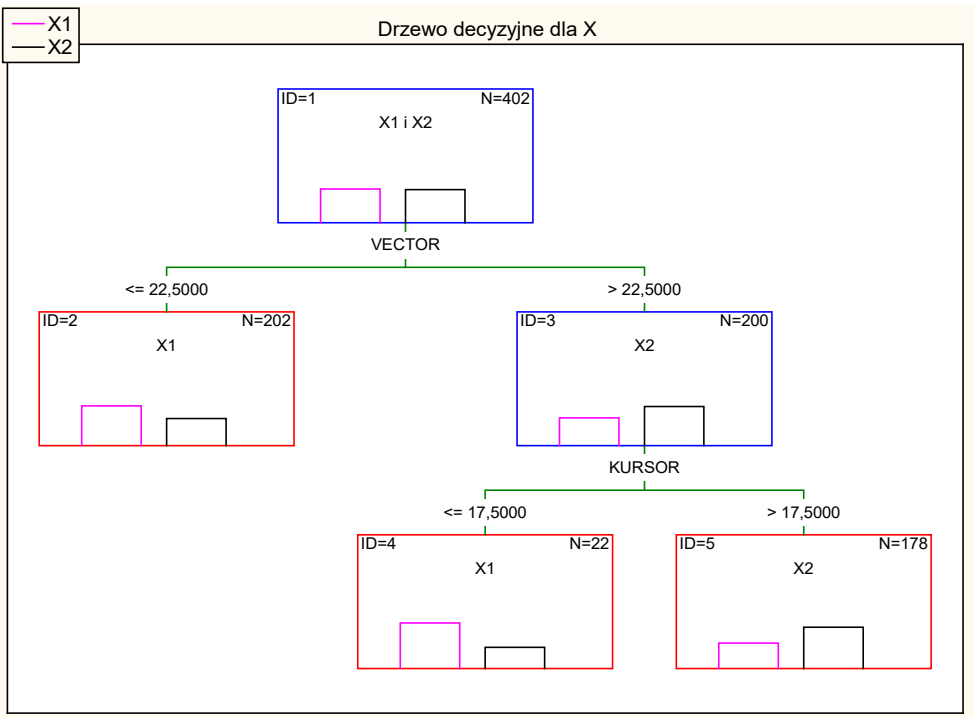


Figure 6.2. Sample Decision Tree Identified Using the Classification and Regression Tree (C&RT) Algorithm

To build a model using the C&RT algorithm, one of the analyzed variables is selected to serve as the dependent (explained) variable, also referred to as the *target*. This may be a nominal, ordinal, or quantitative variable. The target variable is explained using predictors specified by the researcher. The goal of the analysis is to generate a mathematical function that enables the dataset to be divided into homogeneous subgroups—i.e., classes—based on the target variable. The splits are performed iteratively until the model is optimally fitted to the training data.

In the initial stage, the dataset is randomly split into two parts. The standard approach is to allocate 80% of the data for model construction, while the remaining

20% is used to assess model accuracy. The procedure for defining the splitting function continues until the categories of the predictor variables are clearly assigned to categories of the target variable, resulting in internally homogeneous groups. During the analysis, it is also possible to identify which variables from the set of predictors have the greatest influence on the target variable. The outcome of this process is a model which, in the case of the C&RT algorithm, takes the form of a decision tree (Berg, 2008; Nisbet et al., 2009).

The C&RT algorithm is one of many tools used for constructing predictive models. Among alternative methods are the Quinlan algorithm, Automated Neural Networks, Boosted Trees, Support Vector Machines, and MARSplines (Nisbet et al., 2009; Rzechowska, 2004). Each of these methods has specific characteristics; however, the C&RT algorithm has gained popularity due to its simplicity and ability to generate easily interpretable rules.

An example of applying the C&RT algorithm may involve analyzing a variable X, which takes two levels: X_1 and X_2. In a sample of 402 observations, there were 203 observations assigned to category X_1 and 199 observations to category X_2. In this sample, category X_1 was dominant over X_2. The algorithm analyzed the data, searching for optimal split points and ultimately identified them. The first split point was determined for the variable *VECTOR*, where values greater than 22.5 were assigned to category X_2, and values less than or equal to 22.5 were assigned to category X_1. The second split point was based on the variable *KURSOR*, where values above 17.5 were classified as X_2, and values less than or equal to 17.5 were assigned to X_1. The structure of the resulting tree is illustrated in Figure 6.2.

The depth of the tree in the C&RT method depends on both the number of variables in the dataset and the sample size. After the model is constructed, its classification accuracy based on the predictors can be estimated. The results of the accuracy analysis are most often presented as the percentage of correctly classified cases. In the example discussed, the model achieved moderate accuracy of 60.94%, with a testing error of 39.05%. Table 6.2 presents detailed classification results, showing the number of cases correctly and incorrectly assigned to each category based on the predictors.

Table 6.2. Classification Accuracy Results for the Test Model

	Classified as X_1	**Classified as X_2**	**% Correctly Classified**
Observed as X_1	135	68	66.5%
Observed as X_2	89	110	55.27%
			Overall classification accuracy: 60.94%

The data presented in Table 6.2 indicate that the test model exhibited moderate classification accuracy. The analysis shows that, based on the predictors, the algorithm correctly classified observations belonging to category X_1 with an accuracy of

66.5%, while observations from category X_2 were correctly classified with an accuracy of 55.27%. The overall classification accuracy of the model was 60.94%.

A detailed examination of the classification results reveals that, among observations actually belonging to category X_1, the algorithm correctly assigned 135 cases, while 68 were misclassified as X_2. For observations belonging to category X_2, the algorithm correctly assigned 110 cases to X_2, while 89 were incorrectly classified as X_1. These results highlight certain limitations in the model's precision when assigning observations to specific categories, which may stem from the limitations of the predictors or characteristics of the dataset itself.

The structure of the decision tree obtained in this analysis can be presented not only in graphical form but also in a tabular representation detailing the construction of the tree. An example of such a tabular structure is presented in Table 6.3.

Table 6.3. Tree Structure for the C&RT Model

Node No.	Left Branch	Right Branch	Node Size	N Level X_1	N Level X_2	Selected Class	Split Variable	Split Constant
1	2	3	402	203	199	X_1	VECTOR	22.5
2	—	—	202	120	82	X_1	—	—
3	4	5	200	83	117	X_2	KURSOR	17.5
4	—	—	22	15	7	X_1	—	—
5	—	—	178	68	110	X_1	—	—

Columns 'N Level X_1' and 'N Level X_2' indicate the number of cases in a given node assigned to the respective categories of variable X.

This type of table provides detailed information about the structure of the decision tree, including node numbers, branches, node sizes, split points for the predictor variables, and the names of those variables. Such a representation allows for a precise analysis of the tree's structure and the way in which specific variables influence the classification process.

One of the key aspects of decision tree construction is determining the importance of individual predictors—that is, their role in explaining the dependent variable. Predictor importance can be presented in various forms, including tables and histograms that visualize each variable's contribution to the classification process.

Table 6.4 presents the importance of the individual predictors in the construction of the sample tree illustrated in Figure 6.2. Among the four analyzed variables, the highest importance was assigned to the variable *KURSOR*. It was followed by *VECTOR*, *VERSUS*, and *EXIS*. It is worth noting that although *VECTOR* builds the largest branch of the tree, the algorithm identified *KURSOR* as the most important variable. This results from its crucial role in the final data split and classification—though this is not always the case. The importance of predictors depends on the specific characteristics of the dataset and the algorithm applied.

Table 6.4. Predictor Importance

Predictor	Rank Score	Importance
KURSOR	100	1.000000
VECTOR	88	0.877443
EXIS	88	0.876361
VERSUS	73	0.731248
SPECTRUM	54	0.535556
SPACE	48	0.480469
EXTAND	44	0.438891
GRAPH	41	0.405115
POINT	35	0.353698
COMMA	24	0.241018

The predictor importance index can also be presented in the form of a histogram, which facilitates easier interpretation of the results. Figure 6.3 visualizes the importance of the variables in the construction of the decision tree. Taller bars correspond to variables with greater contributions to classification, indicating their key role in the data-splitting process and in explaining the dependent variable.

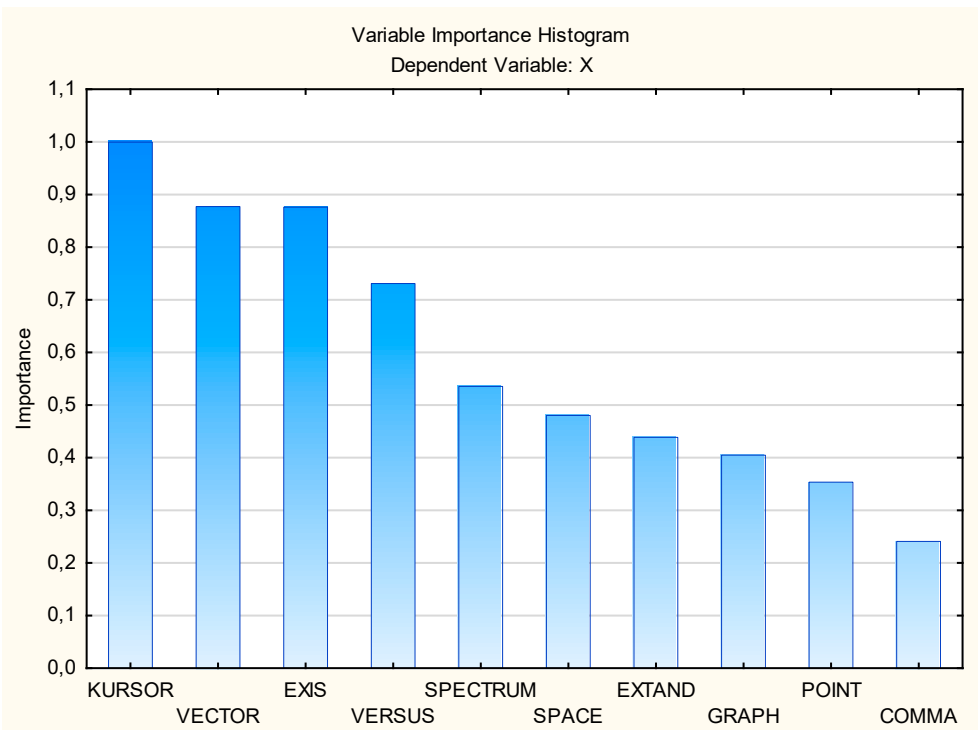


Figure 6.3. Histogram Presenting Predictor Importance

Figure 6.4 illustrates the paths followed by the algorithm while scanning the dataset in order to ultimately assign individuals to groups X_1 and X_2. All paths lead to a point where the variable X takes one of two values—X_1 or X_2. For example, an individual with the highest score on the *VERSUS* scale also displayed a relatively low score on the *EXTAND* variable. In contrast, the person with the highest score on the *VECTRUM* variable also scored highest on the *COMMA* variable and had a moderate score on *KURSOR*.

Tracing the value paths within individual variables allows the algorithm to gather information about the structure and patterns of the data. This enables the algorithm to effectively learn and classify the examined cases. Figure 6.4 presents a complex coordinate network, illustrating the scores achieved by individuals on specific predictors and their final assignment to one of the two dependent groups—X_1 or X_2.

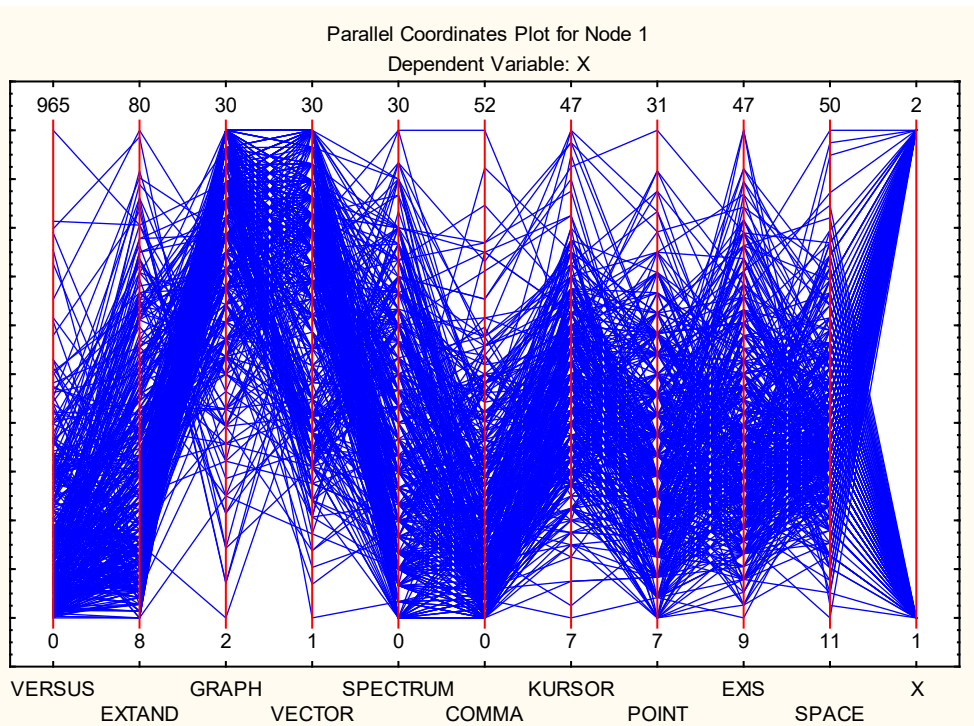


Figure 6.4. Parallel Coordinate Plots in Node 1 for Variable X

The results of participants with respect to each variable and the final classification performed by the algorithm can also be presented in tabular form. This type of representation allows for a detailed analysis, comparison of observed and predicted values, and tracking the process of assigning observations to the appropriate nodes of the decision tree. An example of such data is presented in Table 6.5, which illustrates the values observed and predicted by the algorithm and indicates the assignment of individual observations to specific model nodes.

Table 6.5. Example of Observed and Predicted Values Generated by the Decision Tree Algorithm and Node Assignments

The table shows example results generated by the test model. The numbers indicate the observed and predicted values, classification probabilities for each category, and the terminal node number in the tree structure.

#	Observed Value	Predicted Value	Probability X_1	Probability X_2	Terminal Node
1	X_1	X_1	0.594059	0.405941	2
2	X_2	X_2	0.382022	0.617978	5
3	X_2	X_1	0.594059	0.405941	2
4	X_1	X_1	0.594059	0.405941	2
5	X_2	X_2	0.382022	0.617978	5
6	X_2	X_2	0.382022	0.617978	5
7	X_2	X_2	0.382022	0.617978	5
8	X_2	X_2	0.382022	0.617978	5
9	X_2	X_1	0.594059	0.405941	2
10	X_1	X_1	0.594059	0.405941	2
11	X_1	X_1	0.594059	0.405941	2
12	X_1	X_1	0.594059	0.405941	2

This table presents the observed and predicted values assigned by the decision tree algorithm, as well as the classification probabilities for categories X_1 and X_2. Additionally, it includes information about the tree nodes to which each observation was assigned. This detailed presentation allows for accurate tracking of the classification process and evaluation of the algorithm's performance.

CHAPTER 7

Classification Strategies in Decision Trees with Qualitative Predictors

Classification trees are models in which the dependent variable is qualitative in nature—that is, a nominal or ordinal variable. This variable is explained using predictors that may vary in type, including both qualitative (nominal or ordinal) and quantitative predictors measured on interval or ratio scales. The aim of classification trees is to generate rules for partitioning objects in the dataset in such a way as to maximize the homogeneity of the resulting subgroups with respect to the dependent variable.

Various algorithms are used to construct classification trees, one of the earliest being the algorithm developed by Quinlan (Quinlan, 1993). His work not only initiated the development of this method but also contributed to the popularization of decision trees as an analytical tool (Nisbet et al., 2009).

7.1. Quinlan: A Pioneer of Decision Trees

The first decision tree algorithm, developed by Ross Quinlan, was called ID3 (Iterative Dichotomiser 3) and was introduced in 1986. In 1996, Quinlan presented an improved algorithm named C4.5, which extended and enhanced ID3 (Quinlan, 1993). The ID3 algorithm was based on the analysis of the information content carried by individual attributes through the calculation of their entropy. The attributes that enabled the construction of the most efficient decision tree were then selected. The key criterion for attribute selection was identifying those that resulted in the greatest information gain—i.e., those that most effectively differentiated the objects in the dataset (Quinlan, 1993).

For example, in a dataset containing 300 women and 300 men, an effective split can be achieved using attributes such as hair color, hair length, hairstyle, or presence of facial hair. These attributes may prove useful not only in differentiating the analyzed dataset but also possess generalizability, making them applicable in other datasets of similar nature.

The most valuable attributes are those that allow the dataset to be split into subgroups of similar size while conveying a high amount of information. Such attributes enable a faster and more efficient tree-building process, and the results obtained are more easily generalizable to other datasets. The informational value of attributes is assessed based on their entropy, which is a measure of uncertainty or lack of information. Lower entropy indicates that an attribute carries more information and is more useful in the classification process.

The ID3 algorithm operates in several stages:

1. Calculating entropy for each attribute in the dataset.
2. Selecting the attribute with the lowest entropy, i.e., the one that carries the most information and enables the most balanced split of the dataset.
3. Dividing the dataset into subsets based on the values of the selected attribute.
4. Defining splitting conditions (e.g., what hair length is considered "long" versus "short").

Before discussing entropy calculation in detail, it is important to emphasize that the lower the entropy, the greater the knowledge about the analyzed phenomenon. High entropy, which reflects low information gain, may indicate several issues: poor attribute selection, non-representative sampling, or difficulty in efficiently partitioning the dataset by the algorithm. These issues can lead to flawed models and hinder meaningful generalizations.

For this reason, the selection of key attributes that have the greatest explanatory power for the dependent variable plays a critical role in the decision tree construction process.

Entropy is calculated using Formula 7.1 (Quinlan, 1993):

$$E_j = \frac{n_j^+}{n} I_j^+ + \frac{n_j^-}{n} I_j^- . \tag{7.1}$$

The information content IjI_jIj can be calculated using Formula 7.2:

$$I_j = \sum\nolimits_{i=1}^{n} (-p_i \; log_2 \, p_i) \tag{7.2}$$

where:
E_j – entropy, a measure of uncertainty for attribute *j*,
n_j^+ – number of instances satisfying condition *j*,
n_j^- – number of instances not satisfying condition *j*,
n – total number of instances,
p_i – information content for the respective groups.

Alternatively, entropy can be computed using Formula 7.3 (Quinlan, 1993):

$$E_j = -p_i \ log_2 p_i - (1 - p_i) log_2 (1 - p_i) \tag{7.3}$$

where:
E_j – entropy, a measure of uncertainty,
p_i – probability of a positive instance in the dataset,
$1 - p_i$ – probability of a negative instance in the dataset.

To better understand how the algorithm selects attributes by calculating entropy, let us consider a simple example related to psychotherapy. Ten respondents provided answers regarding the effectiveness of psychotherapy. Among them, six people reported that psychotherapy was effective, three had difficulty assessing its effectiveness, and one person stated that psychotherapy led to deterioration.

The respondents' relationship with their psychotherapist was also examined. These data are presented graphically in Figure 7.1, which facilitates interpretation of the dataset. It is important to emphasize that this visualization serves merely an illustrative purpose and does not depict the output of a decision tree. These data will be used to demonstrate the entropy calculation process—a measure of uncertainty or lack of information in a dataset. Entropy analysis enables the identification of the attribute—whether the **relationship with the psychotherapist** or the **setting of the therapy**—that most significantly reduces uncertainty. The attribute carrying the most information should be selected first in constructing the decision tree.

The preliminary data regarding the relationship with the psychotherapist are as follows:

- Among the six individuals who deemed psychotherapy effective, five reported a good relationship with their psychotherapist, while one described the relationship as difficult (see left branch in Figure 7.1).
- Of the three individuals who were unable to assess the effectiveness of psychotherapy, one had a good relationship, and two described it as difficult (middle branch in Figure 7.1).
- The one person who experienced a deterioration as a result of therapy reported a difficult relationship with the psychotherapist (right branch in Figure 7.1).

These data form the basis for further calculations, such as entropy analysis, which allows for evaluating the extent to which the variable *"relationship with the psychotherapist"* contributes to constructing an effective decision tree.

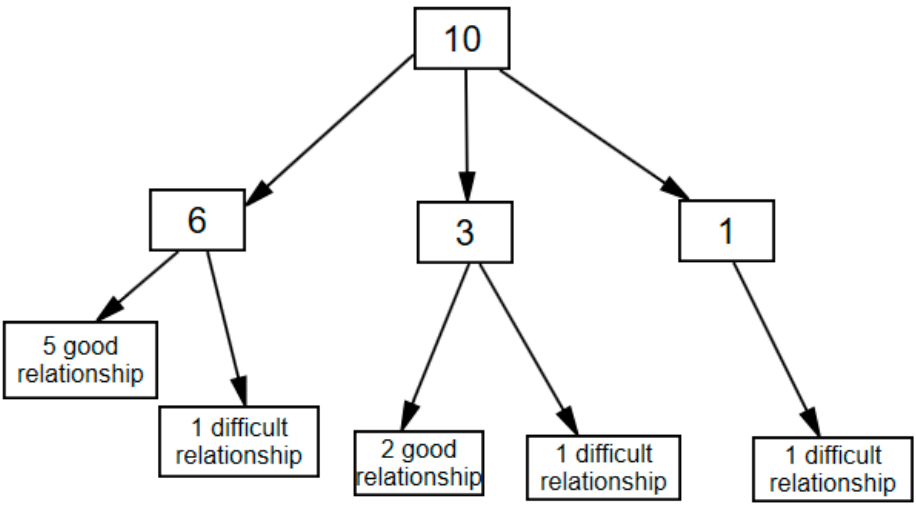


Figure 7.1. Graph presenting the distribution of individuals in the example by good vs. difficult relationship with the psychotherapist

In addition, the setting in which psychotherapy took place was examined. These data are presented in Figure 7.2 as a graphical visualization to support interpretation of the dataset. It should be emphasized that this is only an auxiliary visualization, and final results will be obtained in subsequent stages of the analysis, including through entropy calculation.

The data concerning the setting in which therapy was conducted are as follows:

- Among the six individuals who considered psychotherapy effective, three underwent therapy at a private clinic and three through the NFZ (see the left branch in Figure 7.2).
- Of the three individuals who had difficulty evaluating the effectiveness of psychotherapy, one participated in therapy at a private clinic and two through the NFZ (middle branch in Figure 7.2).
- The only individual who experienced deterioration as a result of therapy participated in therapy at a private clinic (right branch in Figure 7.2).

These data will be used for an analysis aimed at determining whether the variable "setting of psychotherapy" contributes significantly to building an effective decision tree.

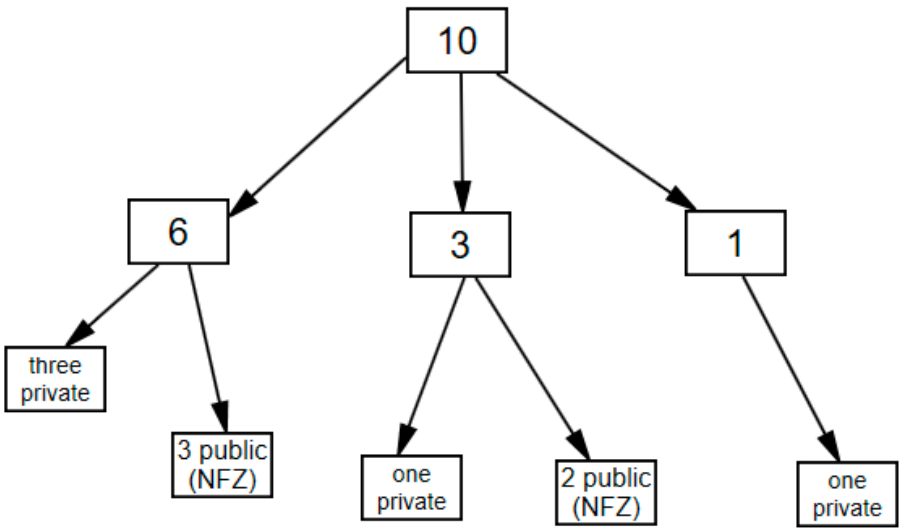


Figure 7.2. Graph showing the distribution of individuals in the example by the setting in which psychotherapy was conducted

The key question is: which attribute – "relationship with the psychotherapist" or "setting of psychotherapy" – carries more information about the effectiveness of psychotherapy? To determine this, entropy calculations were performed for each attribute. Entropy for the variable "relationship with the psychotherapist" was calculated using formula (7.3):

Group of individuals who considered the therapy effective:

- **Good relationship:**

$$-\tfrac{5}{6}\ log_2\tfrac{5}{6} = \left(-0{,}833\right)\cdot\left(-0{,}263\right) = 0{,}2191$$

- **Difficult relationship:**

$$-\tfrac{1}{6}\ log_2\tfrac{1}{6} = \left(-0{,}1666\right)\cdot\left(-0{,}285\right) = 0{,}4307$$

- **Sum:** $\mathbf{0{,}2191 + 0{,}4307 = 0{,}6498}$

Group of individuals who were unable to assess the effectiveness:

- **Good relationship:**

$$-\tfrac{1}{3}\ log_2\tfrac{1}{3} = \left(-0{,}3333\right)\cdot\left(-1{,}585\right) = 0{,}5282$$

- **Difficult relationship:**

$$-\tfrac{2}{3}\ log_2\tfrac{2}{3} = \left(-0{,}6666\right)\cdot\left(-0{,}585\right) = 0{,}3899$$

- **Sum:** $0{,}3899 + 0{,}5282 = 0{,}9181$

Group of individuals who experienced deterioration:
- **Difficult relationship:**

$$-\tfrac{1}{1}\ log_2\tfrac{1}{1} = \big(-1\big)\cdot\ \big(0\big) = 0$$

Total entropy:
- **Total entropy for the attribute**

$$E_1 = \tfrac{6}{10}\cdot 0{,}6498 + \tfrac{3}{10}\cdot 0{,}9181 + \tfrac{1}{10}\cdot 0 = 0{,}3899 + 0{,}2754 + 0 = 0{,}6653$$

Entropy calculations for the attribute "setting of psychotherapy" based on formula 7.2:

Group of individuals who considered the therapy effective:
- **Private clinic:**

$$-\tfrac{3}{6}\ log_2\tfrac{3}{6} = \big(-0{,}5\big)\cdot\ \big(-1\big) = 0{,}5$$

- **NFZ clinic:**

$$-\tfrac{3}{6}\ log_2\tfrac{3}{6} = \big(-0{,}5\big)\cdot\ \big(-1\big) = 0{,}5$$

- **Sum:** $0{,}5 + 0{,}5 = 1$

Group of individuals who were unable to assess the effectiveness of psychotherapy:
- **Private clinic:**

$$-\tfrac{1}{3}\ log_2\tfrac{1}{3} = \big(-0{,}3333\big)\cdot\ \big(-1{,}585\big) = 0{,}5282$$

- **NFZ clinic:**

$$-\tfrac{2}{3}\ log_2\tfrac{2}{3} = \big(-0{,}6666\big)\cdot\ \big(-0{,}585\big) = 0{,}3899$$

- **Sum:** $0{,}5282 + 0{,}3899 = 0{,}9181$

Group of individuals who experienced deterioration:
- **Private clinic:**

$$-\tfrac{1}{1}\ log_2\tfrac{1}{1} = \big(-1\big)\cdot\ \big(0\big) = 0$$

Total entropy:

- **Total entropy for the attribute**

$$E_2 = \frac{6}{10} \cdot 1 + \frac{3}{10} \cdot 0{,}9181 + \frac{1}{10} \cdot 0 = 0{,}6 + 0{,}2754 + 0 = 0{,}8754$$

As illustrated in the example, the attribute "quality of relationship with the psychotherapist" has lower entropy compared to "setting of psychotherapy". This means that the former carries more information about the effectiveness of psychotherapy. The difference between information gain and entropy allows for the selection of the optimal condition or split point for a given attribute. These calculations are performed according to formula (7.4):

(7.4) $$max_j(I - E_i)$$

After calculating entropy, Quinlan's algorithm will select "quality of relationship with the psychotherapist" as the first attribute to split the data, because it carries more information about the effectiveness of psychotherapy than "setting of psychotherapy". The algorithm will split the data into two groups: individuals with a good relationship with the psychotherapist and those with a difficult relationship. This split will create two nodes of the decision tree.

In each group, the algorithm will check whether all individuals belong to one class, for example, whether all individuals in the group with a good relationship considered the therapy effective. If so, the process for that group will end, and the node will be marked as a leaf node, representing a specific class ("effective therapy"). However, if uncertainty remains within the group, meaning that it includes both individuals who found therapy effective and those who were unable to evaluate it, the algorithm will consider the next attribute, in this case "setting of psychotherapy", as a potential condition for further splitting.

If the algorithm selects "setting of psychotherapy" as the next attribute, the data in the mixed group — including both individuals who found therapy effective and those who did not — will be split into individuals attending private therapy and those receiving therapy through the public healthcare system (NFZ). This process will continue until each group becomes homogeneous, meaning that all individuals in the group belong to a single class (e.g., "therapy effective"), or until a point is reached where no attribute significantly reduces entropy.

As a result of this process, a full decision tree will be generated. At the top of the tree will be "quality of relationship with the psychotherapist" as the first attribute, and the following levels will include other attributes, such as "setting of psychotherapy", which allow further refinement of splits. The final leaves of the tree will correspond to the final classes, such as "effective psychotherapy" or "ineffective psychotherapy".

In our case, the algorithm will begin building the tree by splitting the data according to "quality of relationship with the psychotherapist". Further splits within each group will depend on whether uncertainty remains, and if so, "setting of

psychotherapy" may be the next attribute used for splitting. In this way, the algorithm gradually reduces uncertainty in the data and builds a decision tree that enables not only classification but also understanding of the influence of specific attributes on the effectiveness of psychotherapy.

Unfortunately, the ID3 algorithm has some important limitations. The main issue is that it tends to create very detailed decision trees that are overfitted to the data. As a result, not all rules generated by the tree can be generalized, and the final tree structure contains a lot of irrelevant information. In response to these problems, Ross Quinlan developed a new algorithm, C4.5. This algorithm introduced the technique of tree pruning, which removes unnecessary branches, resulting in models that are more general and less prone to overfitting.

Quinlan's algorithm is often used to solve problems related to the classification of qualitative variable levels. Other algorithms are also used in building classification trees, such as Classification & Regression Tree (CART), interaction trees, and boosted trees. These algorithms will be discussed in detail in the following chapters.

7.2. C&RT: A Comprehensive Classification Approach

The C&RT (Classification and Regression Trees) algorithm is an advanced tool for constructing classification trees when the dependent variable is categorical, and regression trees when the dependent variable is continuous. A key distinguishing feature of this algorithm is its specialization in creating binary trees, where each branch splits into two nodes. As a nonparametric method, C&RT does not require assumptions regarding the distribution of variables, making it exceptionally flexible and widely used in various analyses.

The C&RT algorithm was developed by Breiman and his team in 1984 as a method for creating both predictive and descriptive models in the form of decision trees (Kozak & Juszczuk, 2016). The process of tree construction involves the iterative division of the input data set into two groups (nodes) until specified stopping criteria are met. The main stages of the algorithm's operation are discussed below.

Key steps of the C&RT algorithm:

1. Defining criteria for prediction accuracy

At the outset, the algorithm defines criteria for assessing classification quality, such as the ratio of misclassifications to total classifications. It employs the "costs of misclassification" method, with results presented in the form of a classification matrix. This matrix shows how often cases from one class were incorrectly assigned to another class. Classification costs are particularly important when working with imbalanced classes, allowing for the adjustment of weights assigned to different groups.

2. Selecting the Split Point

The algorithm searches for the optimal split point that maximizes classification accuracy. To this end, it uses the Gini index, which measures the "impurity" of a node. The Gini index reaches a value of 0 when a node contains cases from only one class (Kozak & Juszczuk, 2016).

For a binary split, the Gini index is calculated using the formula:

$$G(m) = 1 - \sum_{k=1}^{K} p(k|m)^2 \tag{7.4}$$

where:
E_j *p(k|m)* denotes the probability of class *k* occurring in node *m*,
K is the number of decision classes.

Alternatively, for a binary division, one may use the difference in class probabilities across the nodes, expressed by formula 7.5:

$$\left\{ \frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right] \right\}^2 \tag{7.5}$$

where P_l and P_r are the probabilities of assigning observations to the corresponding nodes.

For continuous variables, the impurity of a node is calculated using formula 7.6:

$$R(t) = \frac{1}{N_w(t)} \sum w_i f_i (y_i - \bar{y}(t))^2 \tag{7.6}$$

where:
$N_x(t)$ is the weighted number of cases in node *t*,
w_i is the weight assigned to case *i*,
f_i is the response frequency,
$\bar{y}(t)$ is the weighted mean in node *t*,
y_i is the value of the variable.

The developers of the C&RT algorithm opted for the Gini index instead of entropy because it better minimizes the risk of uneven data splits. This reduces the likelihood of generating two nodes, one of which contains only a few observations from a single class, while the other is significantly larger and more heterogeneous.

3. Stopping Splits and Tree Pruning

The C&RT algorithm halts data splits based on two main criteria:

a) When all nodes become homogeneous, meaning they contain cases from only one class.
b) When nodes contain too few cases, e.g., fewer than 20, which may prevent further meaningful division.

It is important to note, however, that the algorithm's creators warned against pruning the tree too early, as this could result in overlooking important data structures. Therefore, the C&RT algorithm first builds a full tree, incorporating all possible splits until the data is exhausted. Only after the tree has been fully constructed does pruning occur, with the aim of simplifying the structure without losing essential information (Kozak & Juszczuk, 2016; Steinberg, 2015).

Pruning of the tree built by the C&RT algorithm is performed according to formula 7.7:

$$R_a(T)=R(T)+a|T| \tag{7.7}$$

where:
R(T) denotes the cost of the model on the training set,
T is the number of terminal nodes in the tree,
a is the penalty for model complexity.

If $a = 0$, the tree is maximally expanded, as there is no penalty for additional nodes. As the value of *a* increases, the algorithm seeks to simplify the tree by removing redundant terminal nodes. Thus, the tree is pruned according to the principle of "cost-complexity" (Steinberg, 2015).

Tree pruning involves merging terminal nodes with their "parent", which means that cases assigned to these nodes are consolidated into the previous node. In other words, the removed nodes are omitted, and the tree structure is significantly simplified. For example, if in the full tree nodes 10 and 11 were to follow nodes 2 and 3, respectively, these may be deleted during pruning, and all elements from nodes 10 and 11 reassigned to the parent nodes 2 and 3.

This pruning process maintains a balance between model complexity and its generalizability, thereby helping to avoid the problem of overfitting. As a result, the final tree is more interpretable and better suited for analyzing new data.

4. Determining the Appropriate Tree Size

A decision tree should be complex enough to accurately reflect the phenomenon under investigation, but simple enough to avoid the problem of overfitting. The C&RT algorithm does not stop at building a single tree; rather, it generates a "sequence of nested, pruned trees", allowing for optimal model structuring (Steinberg, 2015).

The performance of the tree is evaluated in two ways:

a) based on new test data,
b) through cross-validation.

During model construction, the algorithm divides the data into a training set, used to construct the tree, and a test set, used to assess prediction quality. The C&RT algorithm accommodates both dependent (target) and independent (predictor) variables, which may be either continuous or discrete. Additionally, the algorithm

handles missing data effectively. It uses available information from the partial data set to predict missing values for other observations.

One of the key advantages of the algorithm is that it does not require class balance within the training set. C&RT automatically accounts for differences in class sizes, weighting the data to balance their impact on the quality of splits. To interpret the quality of splits, one may analyze the class frequencies in a given node relative to their frequencies in the entire dataset. For instance, if the proportion of the positive class in a given node is greater than that of the negative class relative to the entire sample, this may suggest that the node is selectively oriented toward that class.

$$\frac{N_1(node)}{N_1(root)} > \frac{N_0(node)}{N_0(root)} \tag{7.8}$$

where:
N_1 is the number of cases in the positive class,
N_0 is the number of cases in the negative class.

This formula is not a fundamental component of the C&RT algorithm but may serve as an auxiliary method for interpreting class distribution.

Thanks to this, the user of the algorithm can analyze imbalanced classes without the need for prior group equalization. In the case of a qualitative explanatory variable, the process of building a classification tree using the C&RT algorithm involves identifying associations between the levels of the dependent variable and the levels of the explanatory variable. This enables the formulation of rules that reveal relationships between the variables under study, thereby supporting interpretation and decision-making processes.

Example

Figure 7.3 presents an example of a classification tree constructed using the C&RT algorithm, which illustrates the relationship between the child's age group and the upbringing difficulties reported by the parent. The dependent variable in this case was the child's age group, with three levels: 4, 5, and 6 years. The predictor "experienced parental difficulty" was ordinal in nature and included five levels, ranging from "no difficulty" to "very severe difficulty". A detailed description of the study sample and procedures used can be found in Appendix B.

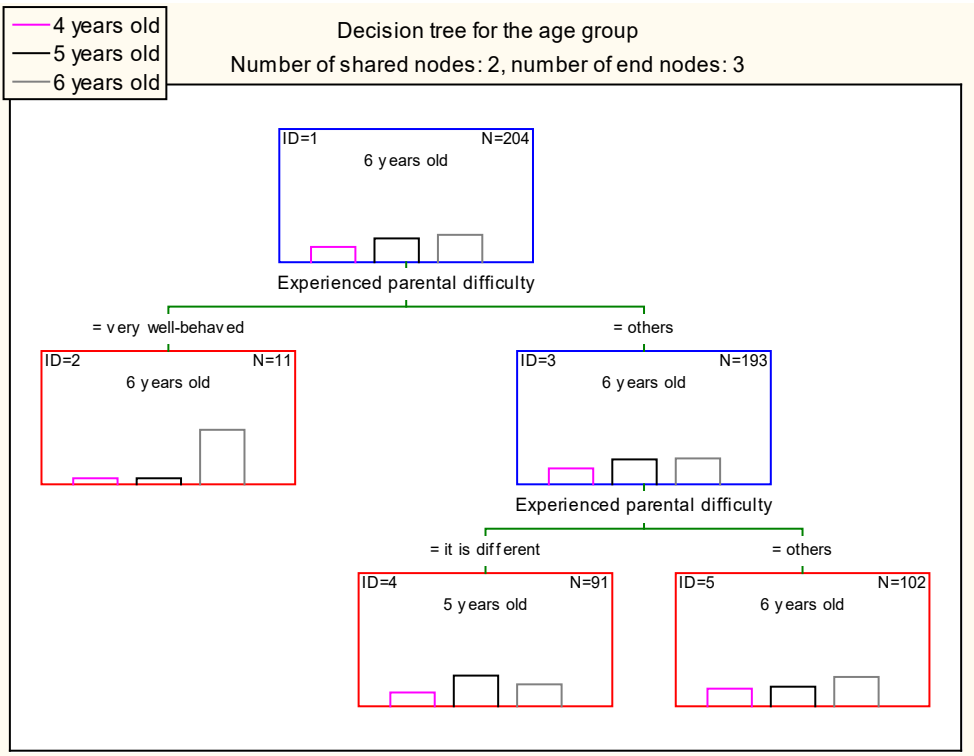


Figure 7.3. Graph of the classification tree built using the C&RT algorithm, illustrating the relationship between the child's age group and the upbringing difficulty experienced by the parent

At the first stage, the algorithm divided the data based on the variable "experienced parental difficulty". As a result of this division, six-year-old children whose parents assessed the upbringing difficulties as "very well-behaved" were assigned to node ID = 2. This group consisted of 11 cases and was nearly homogeneous, meaning that no further splits were necessary. Node ID = 2 thus became a leaf of the decision tree, marked in red in Figure 7.3. The remaining parents who reported other levels of upbringing difficulty were assigned to node ID = 3. This node, marked in blue, required further analysis, as it included cases diverse in terms of both age groups and reported difficulties.

In the next step, the algorithm analyzed group ID = 3, selecting the variable "experienced parental difficulty" at the level "it varies" as the criterion for the next split. As a result of this split, 91 cases were assigned to node ID = 4. This subgroup included 42 parents of five-year-olds, 30 parents of six-year-olds, and 19 parents of four-year-olds. Node ID = 4, marked in red in Figure 7.3, was close to homogeneous, which allowed the division process to conclude for this group. The remaining 102 parents were assigned to node ID = 5, which encompassed a more heterogeneous

group. This group included 27 parents of four-year-olds, 30 parents of five-year-olds, and 45 parents of six-year-olds. Node ID = 5, also marked in red, would require further splitting if the tree-building process were to continue.

The decision tree presented in Figure 7.3 enables the identification of key classification rules. For example, six-year-old children whose parents described them as "very well-behaved" were unambiguously assigned to node ID = 2. Meanwhile, parents of five-year-olds reporting the difficulty "it varies" were mainly assigned to node ID = 4. Group ID = 5, which includes more heterogeneous cases, remains an interesting area for further analysis, as it contains both four-year-old and older children with varying levels of parental difficulties.

Figure 7.3 is supplemented by Table 7.1, which provides detailed information on the number of cases in each node and the criteria used for splitting them. This approach allows for a better understanding of how the C&RT algorithm performs classification and what rules are generated at various stages of tree construction. This analysis is particularly useful for psychologists interested in applying the C&RT algorithm to studies involving qualitative and quantitative variables in empirical psychology.

Table 7.1. Classification Tree Structure

Node ID	Left Branch	Right Branch	Node Size	N (Age 4)	N (Age 5)	N (Age 6)	Selected Class	Splitting Variable	Splitting Threshold	Left Class	Right Class
1	2	3	204	47	73	84	Age 6	Parental Difficulty	very well-behaved		considerable difficulties
2			11	1	1	9	Age 6				
3	4	5	193	46	72	75	Age 6	Parental Difficulty	it varies	some trouble	
4			91	19	42	30	Age 5				
5			102	27	30	45	Age 6				

Based on the rules generated by the C&RT algorithm, the accuracy of predicting a child's membership in a given age group varies by class. For the six-year-old group, the algorithm achieved an accuracy of 64.29%, indicating that most children in this group were correctly classified. For the five-year-old group, the accuracy was slightly lower, at 57.53%, suggesting that the algorithm had greater difficulty predicting membership in this category. For the four-year-old group, however, the algorithm was not able to predict class membership meaningfully, indicating a lack of sufficient distinctions in the data to enable effective classification of this group.

The classification matrix presented in Table 7.2 provides detailed data on the predictions for each age group, offering deeper insight into the algorithm's effectiveness within specific categories.

Table 7.2. Classification Matrix for Individual Age Groups

Observed Group	Predicted Age 4	Predicted Age 5	Predicted Age 6	Total in Group	Best Prediction (%)
Age 4	0	19	28	47	0% (Age 4)
Age 5	0	42	31	73	57.53% (Age 5)
Age 6	0	30	54	84	64.29% (Age 6)
Total	0	91	113	204	

The results of the analysis indicated that parents who described their children as "very well-behaved" were most often the parents of six-year-olds, which the algorithm captured as the dominant rule. In contrast, parents describing their children as "it varies" were mostly parents of five-year-olds. Despite its limitations, these findings partially confirm common developmental patterns observed in five- and six-year-old children, demonstrating the potential of the algorithm in the analysis of upbringing-related data.

In five-year-old children, causal thinking is beginning to develop and reaches a significantly higher level by the age of six (Turner & Helms, 1999). In addition, the rapid motor development typical of five-year-olds results in extremely high activity and dynamism, which may pose challenges for their caregivers. Combined with the emerging causal thinking, five-year-olds often provide their parents with "plenty of excitement". In contrast, six-year-olds, owing to more advanced motor skills and better-developed causal thinking, tend to be more balanced and calm.

The algorithm applied tree pruning after completing the split on the second branch. Had further splits been allowed, the algorithm could have divided groups ID = 5, and possibly also ID = 4, into additional subgroups based on the variable "experienced parental difficulty". However, such further splits would likely not reveal significantly meaningful rules in the remaining part of the dataset. According to the principle of entropy, it is the early branches of the tree that carry the most information, as they are based on the attributes that best differentiate the data. An effective split is one that divides the data into two relatively balanced classes.

Had the tree not been pruned, its accuracy for the analyzed sample might have increased. However, such rules would be specific only to this dataset, greatly limiting their potential for generalization to other populations (Nisbet et al., 2009). For this reason, pruning is essential to maintain a balance between model simplicity and its ability to generalize results.

In conclusion, when building decision trees, particular attention should be paid to the simplicity and informativeness of the attributes used in the model. The C&RT algorithm enables the identification of rules governing the phenomenon under study by revealing relationships between the levels of dependent and explanatory variables.

The analyzed tree was binary in nature, meaning that at each split function, the group was divided into two nodes. In the case of qualitative variables with multiple

levels, such a constraint may affect the efficiency of the model. Therefore, in such cases, it is worth considering algorithms capable of performing simultaneous splits into multiple nodes. These more advanced approaches will be discussed in the following chapter.

7.3. CHAID: Independence Analysis in Classification Trees

The Chi-squared Automatic Interaction Detector (CHAID) algorithm is one of the oldest and most versatile classification algorithms, designed to build non-binary trees, meaning that more than two branches can emerge from a single node. Thanks to this characteristic, CHAID stands out among other algorithms such as C&RT, which restrict the tree structure to binary splits.

CHAID is a nonparametric method, meaning that it does not require the assumptions of variable distribution, such as normality or homogeneity of variance. This makes it particularly useful in empirical data analysis, where such assumptions are often not met. The algorithm relies on statistical tests, such as the chi-squared test (χ^2) of independence, which allows for assessing the relationship between the dependent variable (target) and the predictors.

One of CHAID's key features is its ability to create multi-way split tables that clearly present relationships between variables, especially in situations where both the dependent variable and the predictors have multiple levels. The algorithm merges predictor categories that are statistically similar in terms of their influence on the dependent variable, which helps simplify the tree structure and increase its interpretability.

CHAID was developed by Gordon Kass in 1980 and has since been widely applied in fields such as market research, education, and medicine. It is particularly valued in the analysis of large datasets, where variables have many categories and the aim is to identify key factors influencing the phenomenon under investigation.

Example applications of CHAID include:

- Studying consumer preferences in marketing
- Evaluating the effectiveness of educational programs
- Classifying patients based on medical data

The CHAID algorithm is designed for building classification trees where both the dependent and explanatory variables are measured on ordinal or nominal scales. Thanks to its flexibility and ability to visualize results, it is a tool frequently used in the analysis of both qualitative and quantitative data.

Operation of the CHAID Algorithm

The operation of the CHAID algorithm is based on the iterative partitioning of data using statistical methods for evaluating the relationships between variables. Each step of the process involves constructing a contingency table that compares the levels

of the dependent variable (target) with the levels of potential predictors (attributes). A distinctive feature of the CHAID algorithm is its ability to generate non-binary trees, in which more than two branches may emerge from a single node. This makes the algorithm exceptionally flexible when analyzing data with predictors that have multiple levels. For example, for a predictor with 11 categories, the algorithm may generate three nodes, allowing for a more precise representation of the data structure.

In the first step of the algorithm's operation, a contingency table is created for each predictor. This table allows for the examination of dependencies between the dependent variable and the levels of the predictor, which is crucial for the next stages of analysis. The strength of these dependencies is assessed using the chi-squared test (χ^2), which measures the statistical significance of differences between variable levels. For each predictor, a *p*-value is calculated, representing the probability of a random association.

To correct the *p*-value and account for the risk of errors resulting from multiple comparisons, the algorithm applies the Bonferroni correction. For ordinal predictors, this correction is expressed by the following formula:

$$B_{ordinal} = (c-1)(r-1) \tag{7.9}$$

where *c* is the number of predictor categories and *r* is the number of groups resulting from merging predictor levels. In contrast, for nominal predictors, the Bonferroni correction is more complex and calculated using the following formula:

$$B_{nominal} = \sum_{i=1}^{n-1} (-1)^i \frac{(r-i)^c}{i!(r-i)!} \tag{7.10}$$

Based on the corrected *p*-value, the algorithm selects the predictor with the lowest *p*-value that meets the defined significance threshold (e.g., $p \leq 0.05$). The selected predictor becomes the basis for data partitioning at the current node.

The algorithm further simplifies the tree structure by merging predictor categories deemed statistically similar in terms of their effect on the dependent variable. This reduces the number of predictor levels, resulting in clearer and more interpretable outcomes.

The partitioning process continues iteratively until no predictor meets the significance criterion or the number of cases in the node becomes too small to allow further splitting. The algorithm then terminates, producing a transparent classification tree structure that reflects the relationships between variables and facilitates interpretation of the analysis results. CHAID is particularly effective for multi-level data, where traditional binary algorithms may not offer sufficient precision.

Example of the application of the CHAID algorithm

An example of a classification tree built using the CHAID algorithm, based on qualitative predictors, is presented in Figure 7.4. The algorithm was tasked with analyzing the relationship between the levels of the dependent variable—namely, the age group of children (4, 5, and 6 years)—and the levels of the explanatory variable, which consisted of various kindergartens. The aim of the analysis was to determine whether children in particular age groups were evenly distributed across different kindergartens and, consequently, whether the sample showed variability in terms of kindergarten attendance.

The algorithm was applied to a sample of 204 parents of preschool-aged children. The independent variable—kindergarten—allowed the data to be divided into four subgroups. The first subgroup, consisting of 52 individuals, included children attending Kindergartens A, B, and C. The age structure of this group comprised 26 four-year-olds, 10 five-year-olds, and 16 six-year-olds (see Table 7.3).

The division into subsequent groups was based on differences in the levels of the independent variable. The results of these splits and their details will be presented in the following sections of the analysis, highlighting key relationships between kindergartens and the age of children in the study sample.

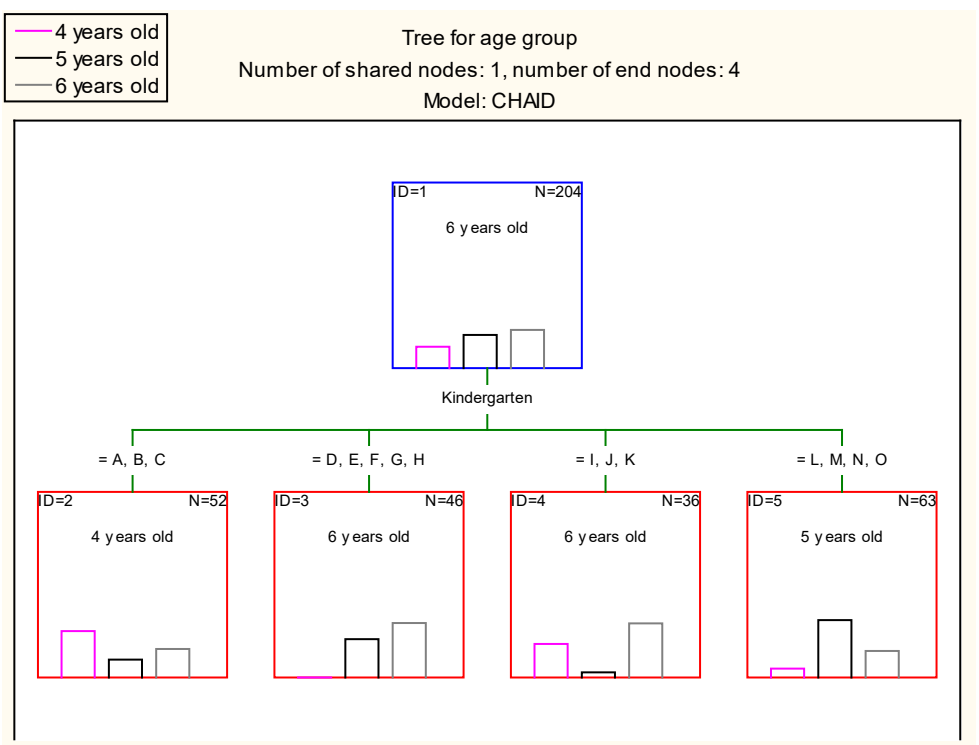


Figure 7.4. Classification tree graph built using the CHAID algorithm, illustrating the relationship between the child's age group and kindergarten attended

The classification tree constructed with the CHAID algorithm demonstrates the data split based on the variable "kindergarten", indicating how children's attendance at specific institutions correlates with their age group. The root node (ID = 1) encompasses the entire dataset, consisting of 204 cases, and is assigned to the dominant class—six-year-old children, who are the most numerous, with 84 cases. The remaining age groups in this node include 47 four-year-olds and 73 five-year-olds.

The first split divides the group into four sub-nodes. Node **ID = 2** includes children attending Kindergartens A, B, and C. This group consists of 52 individuals, with four-year-olds forming the dominant subgroup—26 cases, which is half of the total. It also includes 10 five-year-olds and 16 six-year-olds. Due to the dominance of four-year-olds, this node is classified under the age-4 category.

Node **ID = 3** gathers children attending Kindergartens D, E, F, G, and H. This group contains 46 cases, among which six-year-olds dominate with 27 individuals. The second most numerous group is five-year-olds, with 19 cases. No four-year-olds are present in this subgroup, which distinguishes it from the other nodes. It is therefore classified under the age-6 category.

Node **ID = 4** consists of children from Kindergartens I, J, and K. This group comprises 36 cases, with six-year-olds again forming the majority—21 individuals. The second most represented group is four-year-olds (13 cases), and the least represented group is five-year-olds, with just 2 cases. Given the predominance of six-year-olds, this node is classified under the age-6 category.

The largest sub-node, **ID = 5**, includes children attending Kindergartens L, M, N, and O. This group consists of 63 individuals, with five-year-olds being the dominant group — 39 cases. This subgroup also includes 18 six-year-olds and 6 four-year-olds. Due to the prevalence of five-year-olds, the node is classified under the age-5 category.

To summarize, the partitioning performed by the CHAID algorithm reveals differences in the age distribution of children depending on the kindergarten attended. Node ID = 2 indicates kindergartens where four-year-olds are most common, while nodes ID = 3 and ID = 4 show a dominance of six-year-olds. Node ID = 5, in contrast, contains the highest number of five-year-olds, which may reflect the specific educational programs or admission criteria of the respective institutions. This analysis helps to better understand the population structure of children attending the kindergartens included in the study.

Table 7.3. Classification Matrix for Individual Age Groups

Based on the rules generated by the CHAID algorithm, varying levels of accuracy in predicting a child's membership in a specific age group can be observed. The

Node	Number of Nodes	Node Size	N (Age 4)	N (Age 5)	N (Age 6)	Selected Class	Splitting Variable	Criterion for Child 1	Criterion for Child 2	Criterion for Child 3	Criterion for Child 4
1	4	204	47	73	84	Age 6	Kindergarten	A, B, C	D, E, F, G, H	I, J, K	L, M, N, O
2		52	26	10	16	Age 4					
3		46	0	19	27	Age 6					
4		36	13	2	21	Age 6					
5		63	6	39	18	Age 5					

algorithm achieved 57.78% accuracy in classifying children into the four-year-old group, indicating moderate effectiveness in this category. For five-year-olds, the algorithm's accuracy was slightly lower at 55.71%, suggesting greater difficulty in predicting membership in that group. The highest classification accuracy was achieved for six-year-olds, at 58.54%, making this the best-classified category by the model.

In summary, the CHAID algorithm shows slightly better performance in classifying older children, while its ability to correctly classify four- and five-year-olds is slightly weaker. These results may stem from differences in the data structure, such as uneven sample sizes across age groups or similarities in predictor levels among younger children. Details regarding classification performance are presented in the Classification Matrix (Table 7.4), which shows the distribution of correct and incorrect predictions for each of the analyzed age groups.

Table 7.4. Classification matrix for individual age groups

Observed Group	Predicted Age 4	Predicted Age 5	Predicted Age 6	Total in Group	Best Prediction (%)
Age 4	26	6	13	45	57.78% (Age 4)
Age 5	10	39	21	70	55.71% (Age 5)
Age 6	16	18	48	82	58.54% (Age 6)
Total	52	63	82	197	

The CHAID algorithm successfully linked the levels of the variable "child's age" with the levels of the variable "kindergarten", identifying institutions that had no four-year-olds, as well as those from which the largest number of parents of six-, four-, and five-year-olds originated. With just one splitting function, the results of the algorithm were highly transparent — an especially valuable feature in data analyses involving dependent variables with multiple levels.

Moreover, the CHAID algorithm significantly simplified the structure by merging predictor levels within specific nodes. This grouping was made possible through the analysis of shared associations between predictor levels and the levels of the dependent variable. This process not only streamlined the structure of the tree but also facilitated interpretation of the results, enabling more intuitive conclusions.

The examples of CHAID application presented here primarily illustrate the relationships between variable levels, referred to as main effects. However, it is worth emphasizing that algorithms like CHAID also allow for the analysis of interaction effects, enabling a more detailed understanding of complex relationships among variables. For this reason, the next chapter will focus on interaction-based algorithms that emphasize the relationships between the levels of both explanatory and dependent qualitative variables.

7.4. Interactions in Classification Tree Modeling

The phenomenon under investigation, in addition to main effects, also requires the inclusion of interaction effects, which often play a crucial role in data analysis (Aranowska & Rytel, 2010). Regardless of whether main effects have been identified, it is always advisable to control for interaction effects, as interactions may occur independently of the presence of main effects. Interaction effects arise when the combination of two variable levels determines changes in the dependent variable. For example, different combinations of the levels of factors A and B may influence the dependent variable in specific ways, clearly indicating the presence of an interaction effect.

Let us consider a situation where the gender of the parent (being a mother or a father) alone does not affect the level of demand for obedience from the child. However, the combination of the parent's gender with the child's behavior ("difficult" or "well-behaved") may determine changes in this variable. In such a case, we are dealing with an interaction effect. For instance, being the father of a "well-behaved" child may be associated with a higher level of demand for obedience, whereas being the mother of such a child may relate to a lower level. The opposite pattern may occur for "difficult" children: fathers of these children may show lower levels of demand for obedience, while mothers may show higher levels.

To illustrate this effect, a sample described in Appendix B was analyzed. The sample included 204 individuals, evenly divided into four groups: 51 mothers of "well-behaved" children, 51 fathers of "well-behaved" children, 51 mothers of "difficult" children, and 51 fathers of "difficult" children. Parents answered questions regarding the level of obedience they expected from their children. In this example, the dependent variable (Y) was the level of demand for obedience, while factors α and β referred to the child's behavior ("well-behaved" or "difficult") and the parent's gender (father or mother), respectively. These two-level factors divided the entire sample into four subgroups. The values of the dependent variable for each of these groups are presented in Table 7.5.

Table 7.5. Values of the variable 'demand for obedience' depending on the child's behavior and parent's gender

$\alpha \setminus \beta$	Father	Mother	Marginal Mean
Difficult	40.76	42.08	41.42
Well-behaved	41.12	39.96	40.54
Marginal Mean	40.94	41.02	

As shown, the marginal means do not indicate any main effects for the parent's gender ($F(1,200) = 0.01$; $p = 0.916$), nor for the child's behavior ($F(1,200) = 1.40$; $p = 0.238$). In other words, it cannot be concluded that parents of "well-behaved" children demand more obedience than parents of "difficult" children. Likewise, there is no evidence that fathers or mothers demand more obedience from their children, as no main effect for the parent's gender was detected. However, an interaction effect between the levels of factors α and β in their influence on the dependent variable—demand for obedience—was found ($F(1,200) = 2.75$; $p = 0.049$, one-tailed significance). Figure 7.5 illustrates the mean values for the four groups, allowing these effects to be visualized more clearly.

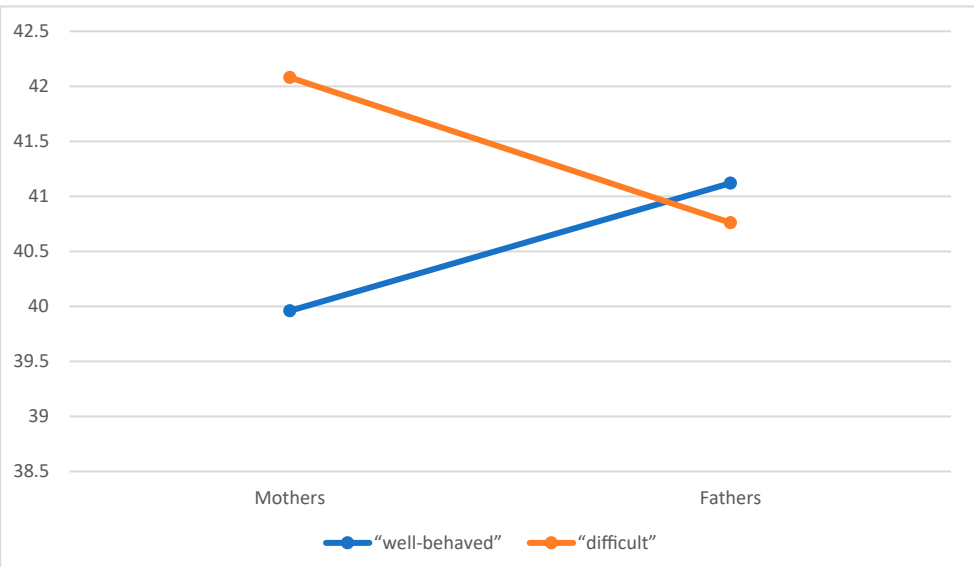


Figure 7.5. Graph illustrating the interaction effect of factors α and β on the dependent variable: demand for obedience from the child

The highest mean levels of demand for obedience were observed in two groups: among mothers of "difficult" children (mean = 42.08) and fathers of "well-behaved" children (mean = 41.12). The lowest means were recorded in the groups of mothers of "well-behaved" children (mean = 39.96) and fathers of "difficult" children (mean = 40.76). The interaction effect between the parent's gender and the child's

behavior proved to be statistically significant, indicating that the observed differences are not random but stem from actual relationships within the studied groups. These results were derived from psychological research and published in the journal *Psychologia Rozwojowa* (Szymańska, 2009).

However, one can imagine a situation in which both interaction effects and main effects occur. To illustrate this, a fictional example is provided. Let us assume that a main effect for factor β (parent's gender) and an interaction effect for factors α and β have been detected. Table 7.6 presents fictional values for groups under this scenario.

Table 7.6. Fictional values for groups assuming the presence of a main effect and interaction

$\alpha \setminus \beta$	Father	Mother	Marginal Mean
Difficult	70	60	65
Well-behaved	90	50	70
Marginal Mean	80	55	67.5

In this case, fathers demand more obedience from children than mothers (fathers' mean = 80; mothers' mean = 55). An interaction effect between the levels of factors α and β is also evident. Fathers of "well-behaved" children demand the most obedience, while mothers of "well-behaved" children demand the least. Parents of "difficult" children show intermediate values (fathers' mean = 70; mothers' mean = 60). The significant main effect for factor α (child's behavior) may be explained by the interaction effect—the fathers of "well-behaved" children exhibit such a high average level of demand for obedience that they raise the mean for the entire group.

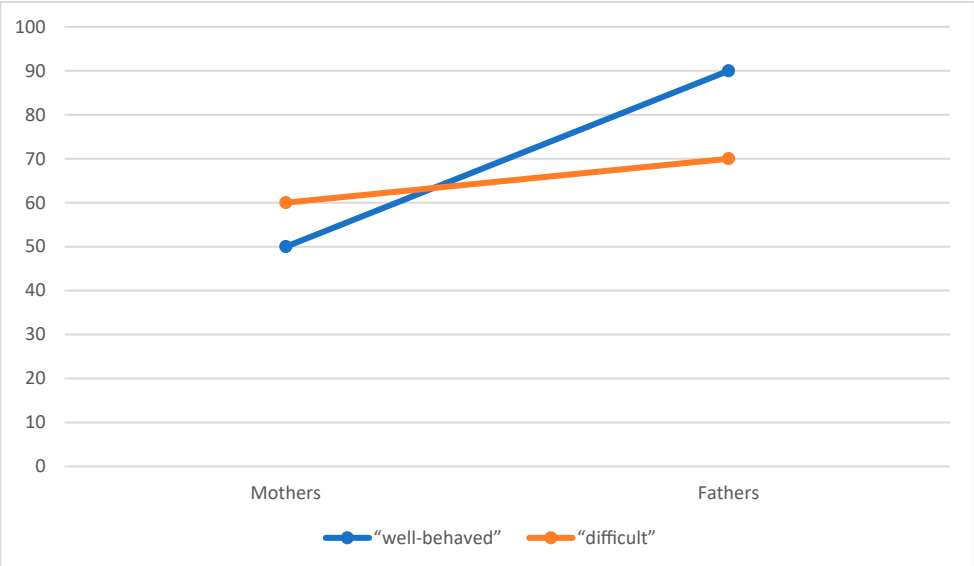


Figure 7.6. Graph illustrating interaction effects and the presence of main effects

Figure 7.6 presents a chart showing the interaction effects between the levels of factors α and β in relation to the dependent variable—demand for obedience from the child. It is evident that the lines connecting the group means for factors α and β intersect, indicating the presence of an interaction effect. This time, the lines representing the means for fathers and mothers are substantially distant from one another (fathers' mean = 80; mothers' mean = 55), which suggests a main effect for factor β.

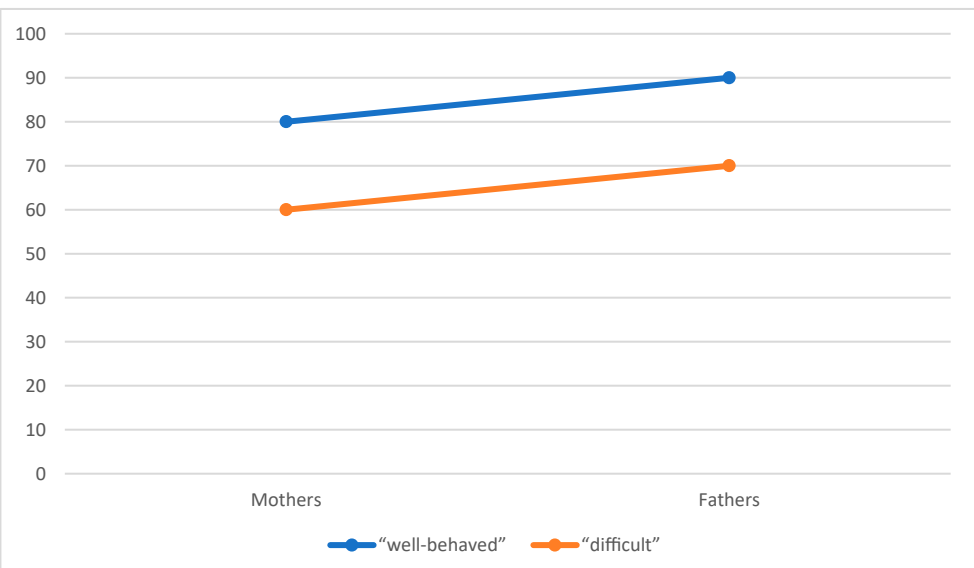


Figure 7.7. Graph illustrating the absence of interaction and the presence of main effects

The lines in the graph do not intersect, which indicates the absence of an interaction effect while simultaneously showing the presence of main effects. Due to possible data configurations that may explain, for instance, an apparent main effect (as in the second example), it is recommended to test for interaction effects in all cases, regardless of whether main effects are present (Aranowska & Rytel, 2010; Kinnear & Gray, 2008).

Conducting research using data mining methods also allows for the testing of interaction effects. Therefore, it is important to consider the possibility of interactions between attributes when explaining the dependent variable. To test interaction effects, so-called interaction trees can be used. This chapter will now present an interaction tree constructed using the C&RT algorithm, as well as one built using the CHAID algorithm.

7.4.1. Application of C&RT in Interactive Models

The previous chapter discussed the interaction between demanding obedience from the child and two factors: the child's behavior in preschool (α) and the parent's gender (β). The dependent variable was numerical, and the factors were qualitative and binary. In this chapter, the child's behavior in preschool (α) is adopted as the dependent variable, while the explanatory factors are the parent's gender (β) and the parenting difficulty (γ) experienced by the parent. The developed classification tree model was built based on qualitative predictors, which means that both the dependent and independent variables are qualitative in nature.

The Classification and Regression Tree (C&RT) algorithm was applied to explain the variable "educator's opinion", which had two levels: "well-behaved child" and "difficult child". The predictors were parenting difficulty (γ), which included five levels: "very well-behaved", "rather well-behaved", "it varies", "I have some trouble with them", and "considerable parenting difficulties", and the parent's gender (β) with two levels: "mother" and "father". The algorithm constructed an interaction tree that controlled for the influence of the levels of these factors on the dependent variable, that is, the educator's opinion. The tree model is presented in Figure 7.8.

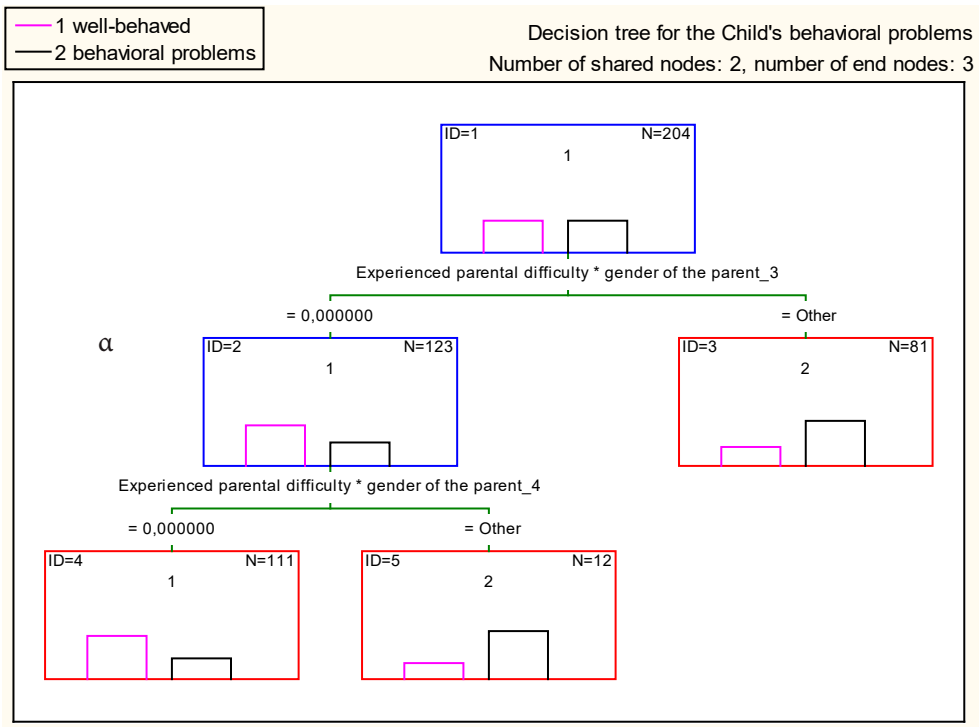


Figure 7.8. Tree graph illustrating the interaction between levels of the variables "experienced parental difficulty" and "parent's gender" in explaining the teacher's assessment

The structure of the classification tree is presented in Table 7.8. This table provides a tabular representation of the information depicted in the decision tree graph, including node number, branch numbers, node size, class membership, and details of variable splits.

Table 7.8. Structure of the classification tree

Node No.	Left Branch	Right Branch	Node Size	Class N_1	Class N_2	Selected Class	Split Variable
1	2	3	204	102	102	1	Experienced parental difficulty * Parent's gender_3
2	4	5	123	78	45	1	Experienced parental difficulty * Parent's gender_4
4			111	75	36	1	—
5			12	3	9	2	—
3			81	24	57	2	—

Table 7.9 presents column labels essential for understanding the structure and interpretation of the decision tree graph. In the label column, one can see that "experienced difficulty" in interaction with "parent's gender" corresponded to different levels of these variables. For example, the first level of parental difficulty ("very well-behaved child") was paired with the first level of gender ("mother"). The next level, "rather well-behaved", was also associated with "mother". The third level of difficulty ("it varies") and the fourth level ("I have some trouble") also interacted with being a mother.

In **Figure 7.8** of the decision tree, it is evident that the first split was based on the interaction between the third level of parental difficulty ("it varies") and the parent's gender ("mother"). Women who described their interactions with the child as "it varies" formed a group in which the majority of children were classified by preschool teachers as "difficult" (57 children). This group also included 24 "well-behaved" children. The total number of cases in this node was 81 (see Node ID 3 in Figure 7.8 and Table 7.8).

For the remaining 123 individuals, the algorithm performed an additional splitting function. This time, it identified a group based on the fourth level of experienced parental difficulty ("I have some trouble with them") in interaction with the parent's gender ("mother"). This group consisted of 12 individuals, most of whom were mothers of "difficult" children (9 cases), while only 3 mothers described their children as "well-behaved" (see Node ID 5, Figure 7.8).

The interaction between experienced parental difficulty and parent's gender indicates that the way in which parents perceive their children varies depending on both their gender and the level of reported difficulty. The results show that mothers are more likely to perceive their children as "difficult" at the third ("it varies") and fourth ("I have some trouble with them") levels of parental difficulty, whereas this effect was not statistically significant for fathers. This may suggest that mothers are more inclined to notice and classify their children's behavior as problematic in everyday interactions. The absence of significant interactions for fathers suggests that they perceive parental difficulties in a more uniform manner, regardless of the difficulty level.

These differences may stem from the distinct social roles that mothers and fathers typically assume in child-rearing. Mothers are often more engaged in daily caregiving challenges, which may heighten their sensitivity to children's behavior in more nuanced ways. These findings also suggest the need to tailor parenting support strategies based on the parent's gender, to more effectively address both "difficult" and "well-behaved" child behaviors.

Table 7.8 and the decision tree graph in Figure 7.8 provide detailed information on the number of cases in each node and the interactions identified between qualitative variables.

Table 7.9. Column labels

Label	Column	Variable	Level of variable	Relative level of variable	Level of variable	Relative level of variable
Experienced parental difficulty	1	Experienced parental difficulty	1	5		
Experienced parental difficulty	2	Experienced parental difficulty	2	5		
Experienced parental difficulty	3	Experienced parental difficulty	3	5		
Experienced parental difficulty	4	Experienced parental difficulty	4	5		
Parent's gender	5	Parent's gender	1	2		
Experienced parental difficulty * Parent's gender_1	6	Experienced parental difficulty	1	5	Parent's gender	1
Experienced parental difficulty * Parent's gender_2	7	Experienced parental difficulty	2	5	Parent's gender	1
Experienced parental difficulty * Parent's gender_3	8	Experienced parental difficulty	3	5	Parent's gender	1
Experienced parental difficulty * Parent's gender_4	9	Experienced parental difficulty	4	5	Parent's gender	1

The algorithm achieved an accuracy of 73.53% in classifying "well-behaved" children and 64.71% in classifying "difficult" children, indicating a moderate effectiveness in predicting children's behavior based on the analyzed factors. Detailed results are presented in Table 7.10.

Out of the 102 children classified as "well-behaved", the algorithm correctly classified 75, while 27 were misclassified, despite incorporating interactions related to how the parent (mother) described their relationship with the child. In the case of "difficult" children, the algorithm correctly classified 66 cases, while 36 were misclassified based on the same criteria.

Notably, the algorithm did not identify significant interaction effects for the "father" level, even though the sample was balanced with respect to parent gender. This may suggest that interaction effects related to parenting difficulties are more pronounced in the context of mothers, which warrants further analysis.

Table 7.10. Classification matrix for the "well-behaved child" and "difficult child" classes

Observed group	Predicted 1	Predicted 2	Total in group	Best prediction (%)
1	75	27	102	73.53% (1)
2	36	66	102	64.71% (2)
Total	111	93	204	

The decision tree model demonstrated that the third level of experienced parenting difficulty ("it varies") and its interactions with the parent's gender—especially in the case of mothers—played a key role in classification. Other levels of parenting difficulty, such as "I have some trouble with them" and "rather well-behaved", also had a significant influence on the analysis outcome, though to a lesser extent than level three.

The algorithm achieved moderate accuracy, correctly classifying 64.71% of "difficult" children and 73.53% of "well-behaved" children. These results highlight the importance of analyzing the interaction between parenting difficulties and the parent's gender in explaining a teacher's perception of child behavior.

7.4.2. CHAID in Interaction Analysis

In this study, interactions between the child's age (λ) and the preschool (δ) attended were analyzed to determine their impact on the teacher's opinion about the child's behavior ("well-behaved" or "difficult"). The CHAID algorithm was used to identify from which preschools and age groups the largest number of "difficult" children came, and to verify whether the distribution was uniform. The aim of the analysis was to build a classification decision tree using qualitative predictors.

The dependent variable was the teacher's opinion of the child, with two levels: "well-behaved child" and "difficult child". The factor "child's age" (λ) was a qualitative variable with three levels: four-year-olds, five-year-olds, and six-year-olds. The "preschool" factor (δ) included 14 levels, representing the preschools that participated in the study. The CHAID algorithm constructed a decision tree, and the results are shown in Figure 7.9.

The first branch of the tree was built based on the interaction between the preschool and the group of four-year-old children. The algorithm isolated 10 "difficult" children from a particular preschool (Node ID = 3). The second branch addressed the interaction of the preschool with five-year-old children, identifying 6 "difficult" children. The final branch once again referred to five-year-olds, emphasizing their affiliation with a different preschool. These results suggest that certain preschools may be more conducive to "difficult" behavior among children in specific age groups.

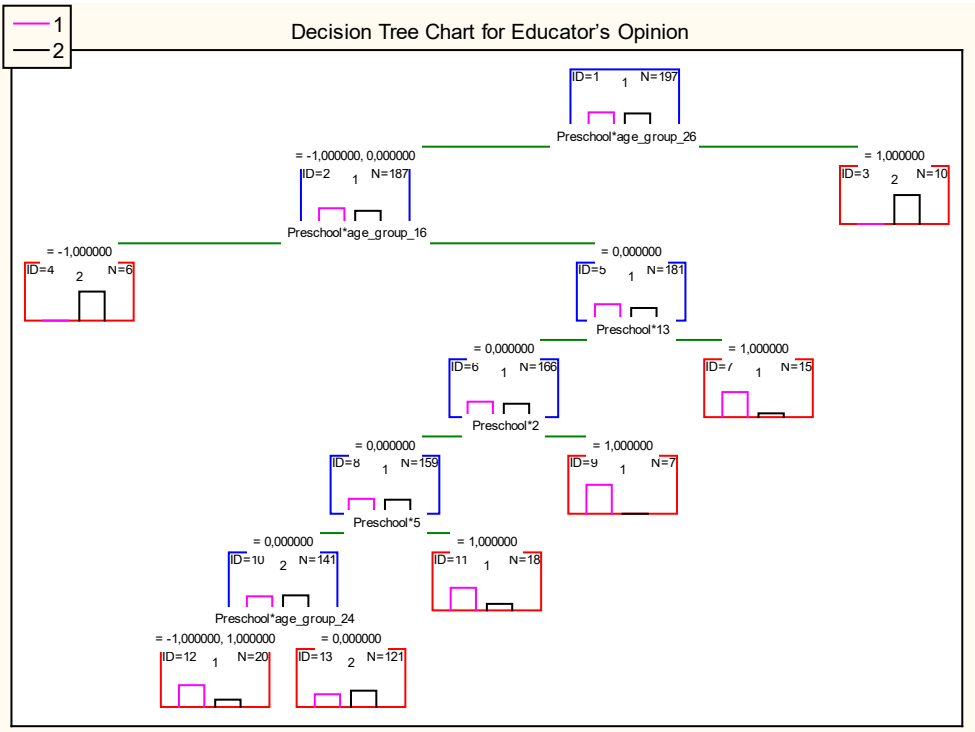


Figure 7.9. Interaction tree model constructed using the CHAID algorithm

Table 7.11 presents the structure of the interaction tree built by the CHAID algorithm, taking into account the interaction effects between kindergartens and children's age groups.

Table 7.11. Structure of the interaction tree constructed using the CHAID algorithm

Node No.	Node size	Class 1 (well-behaved)	Class 2 (difficult)	Selected class	Split variable	Criterion for child 1	Criterion for child 2
1	197	102	95	Well-behaved	Kindergarten*age_group_26	–1, 0	1
2	187	102	85	Well-behaved	Kindergarten*age_group_16	–1	0
4	6	0	6	Difficult			
5	181	102	79	Well-behaved	Kindergarten_13	0	1
6	166	89	77	Well-behaved	Kindergarten_2	0	1
8	159	82	77	Well-behaved	Kindergarten_5	0	1

The interaction between the kindergarten and the child's age indicates that specific conditions prevailing in kindergartens may influence the classification of children's behaviour as "difficult". This may be related to differences in upbringing approaches, group sizes, or teachers' competencies. The algorithm achieved the highest effectiveness in classifying "difficult" children (88.42%), which suggests that these children exhibited a more distinct behavioural profile. The classification of "well-behaved" children was less accurate (48.04%), which may indicate greater variability in their behaviour and more difficulty in identifying patterns based on the data (Table 7.12).

Table 7.12. Classification matrix for the interaction model built using the CHAID algorithm

Observed group	Predicted well-behaved	Predicted difficult	Total in group	Best prediction (%)
Well-behaved	49	53	102	48.04%
Difficult	11	84	95	88.42%
Total	60	137	197	

The classification matrix can also be visualised in the form of a bar chart, as shown in Figure 7.10. Based on this figure, it can be observed that the algorithm classified slightly more children from the "well-behaved" group as belonging to the "difficult" group. In contrast, in the case of children who actually belonged to the "difficult" group, the vast majority were correctly classified. It is therefore evident that the CHAID algorithm handled the classification of "difficult" children effectively, while the classification of "well-behaved" children was less accurate.

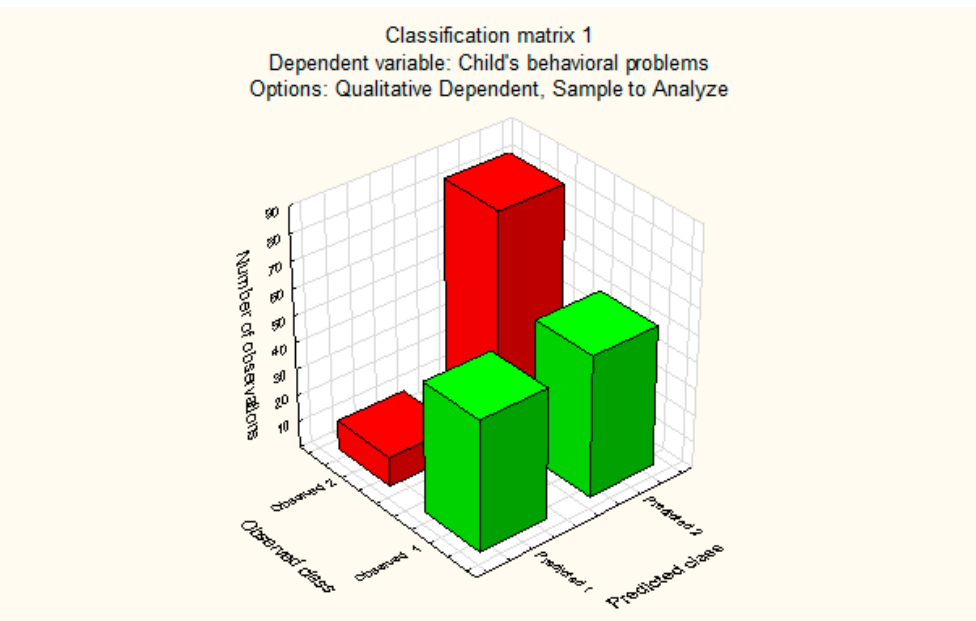


Figure 7.10. Classification into the groups of "well-behaved" and "difficult" children based on the CHAID algorithm solution

The results of the CHAID analysis highlight the importance of tailoring **educational strategies** to the specifics of age groups and kindergarten conditions. Interactions between the child's age and the kindergarten may indicate areas requiring improvement, such as group size, methods of working with children, or teacher support. The findings also suggest that a more detailed analysis of the behaviour of "well-behaved" children could help improve their classification and identify factors that foster positive behaviours.

CHAPTER 8

Quantification of Variables in Decision Trees with Quantitative Predictors

Classification trees with quantitative predictors are models in which the dependent variables (also referred to as target variables) take qualitative values, i.e. values expressed on nominal or ordinal scales, whereas the predictors (explanatory variables) are quantitative in nature, expressed on interval or ratio scales.

These models offer an alternative to discriminant analysis (Kinnear & Gray, 2008), providing several significant advantages:

a) They are non-parametric methods, which means they are resistant to assumptions regarding distribution normality and variance homogeneity.
b) The tree-building process involves the iterative splitting of data using predictors until maximum homogeneity within subgroups is achieved.
c) The transparency of the splitting rules allows for an easy understanding of the relationships between the dependent and explanatory variables.
d) Tree-generating algorithms make it possible to create predictive rules that can be successfully applied to other datasets.

Thanks to these features, classification trees with quantitative predictors are particularly useful in data analysis where it is important to capture interactions between variables and to simplify complex relationships within the data.

This chapter discusses three approaches to building classification trees with quantitative predictors: the binary C&RT algorithm, the non-binary CHAID algorithm, and interaction trees, along with application examples and interpretation of their results.

8.1. C&RT for Quantitative Predictors: Methodology and Applications

The C&RT algorithm, used for building classification trees with quantitative predictors, can be described as a non-parametric counterpart to logistic regression or discriminant analysis. In such models, the dependent (target) variable is measured on a qualitative scale (nominal or ordinal), while the predictors are quantitative in nature and are expressed on interval or ratio scales.

This method enables the classification of the target variable's objects based on the numerical values of the predictors. The algorithm identifies specific splitting points for each explanatory variable, allowing the extraction of homogeneous groups in terms of the target variable. This chapter presents the application of the binary C&RT algorithm, where the target variable was *parental difficulty*, and the predictors were four temperamental traits of the children.

In the analysed sample, parental difficulties were standardised. Scores above 0.5 standard deviations were assigned to the "difficulty" group, and scores below –0.5 to the "no difficulty" group. Observations with scores ranging from –0.5 to 0.5 were excluded from the analysis. As a result, the dataset included 145 observations.

The aim of the analysis was to identify the temperamental traits that most significantly differentiated between parents experiencing parental difficulties and those who did not. The constructed decision tree demonstrated a high classification accuracy of 86.49%. Table 8.1 presents the detailed classification results, including both correctly and incorrectly classified cases.

The analysis results indicate a high level of model accuracy. The algorithm correctly classified 81.61% of parents experiencing difficulties, and 91.38% of parents in the "no difficulty" group.

Table 8.1. Classification results for the tree

	Classified as difficulty	Classified as no difficulty	% correctly classified
Observed as difficulty	71	16	81.61%
Observed as no difficulty	5	53	91.38%
Overall accuracy			86.49% average

The tree showed high accuracy for both high and low levels of difficulty in the parent–child relationship.

Table 8.2. Predictor importance

	Predictor rank	Importance
Temperamental trait 1	100	1.000000
Temperamental trait 2	95	0.954458
Temperamental trait 3	77	0.777121
Temperamental trait 4	41	0.410360

Table 8.2 presents the importance of individual predictors within the model. The most significant variable was *temperamental trait 1*, followed by *traits 2*, *3*, and *4*. Figure 8.1 illustrates the structure of the decision tree, including the hierarchy and the split points identified by the algorithm.

The first differentiating attribute forming the main branch of the tree was *temperamental trait 2*. The algorithm identified a split point at the level of 15.5. If the value of this trait was greater than 15.5, the dependent variable was assigned to the "high difficulty" category (Group 2). Values less than or equal to 15.5 were associated with the "low difficulty" category (Group 1). The results indicate that the greater the intensity of *temperamental trait 2*, the higher the level of difficulty experienced by parents in the relationship with their child.

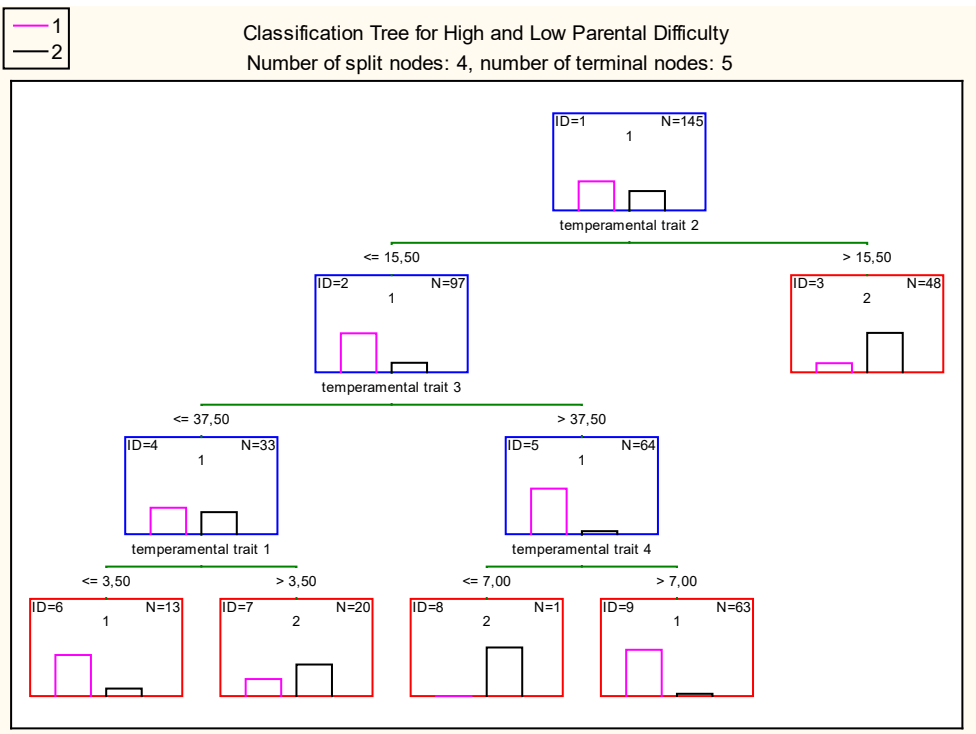


Figure 8.1. Decision tree for parental difficulty in upbringing situations

The second split point was the value of 37.5 for *temperamental trait 3*. If its value exceeded 37.5, nearly all cases were classified as "low difficulty" (Group 1). This group included 64 individuals who experienced low levels of difficulty (*ID5*, *N = 64*). When the value of *temperamental trait 3* was less than or equal to 37.5, the classification became less clear, with the number of cases experiencing difficulty and no difficulty being nearly equal (*ID4*, *N = 33*).

The third key predictor was *temperamental trait 1*. The algorithm identified a group of 20 individuals (*ID7*, *N = 20*) in which the majority experienced difficulties

in the parent–child relationship (Group 2). These individuals had values of *temperamental trait 1* greater than 3.5. For values less than or equal to 3.5, almost all parents experienced low levels of difficulty (Group 1) (*ID6*, *N = 13*).

The fourth split point was *temperamental trait 4*. Parents whose values on this trait exceeded 7 most often experienced low levels of difficulty (*N = 63*, Group 1). In contrast, values less than or equal to 7 were clearly associated with high levels of difficulty in the parent–child relationship (Group 2).

The analysis revealed characteristic patterns in the data. The largest group of parents (*N = 63*, *ID9*) who experienced low difficulty had children characterised by high levels of *temperamental trait 4*, high levels of *trait 3*, and low levels of *trait 1*. Conversely, the largest group of parents experiencing difficulties (*N = 48*, *ID3*) was associated with high values of *temperamental trait 1*.

The decision tree allows for a detailed characterisation of the identified groups of parents, revealing the relationships between children's temperamental traits and the difficulties experienced by parents.

8.2. CHAID: Quantitative Classification Techniques

The CHAID algorithm is used in constructing non-binary classification trees with quantitative predictors, which means that the nodes of such trees may branch into more than two branches. In this type of model, the dependent variable is qualitative, while the predictors are measured on quantitative scales. This chapter presents an example of an analysis using the CHAID algorithm.

The study focused on a quantitative variable referred to as *warm directiveness*, whose values were used to predict the qualitative variable *teacher's opinion*. This variable had two levels: "well-behaved child" and "difficult child". The CHAID algorithm analysed the data, identifying three key split points for the *warm directiveness* variable, which enabled a classification accuracy of 82.35% for "well-behaved" children and 49.02% for "difficult" children (Table 8.4).

Table 8.3. Classification matrix of objects performed by the CHAID algorithm for the variable *teacher's opinion*

Observed	Classified as well-behaved	Classified as difficult	% correctly classified
Well-behaved children	84	18	82.35%
Difficult children	52	50	49.02%
Overall	136	68	66.67%

Figure 8.2 presents a graphical representation of the classification tree, and Table 8.4 presents the detailed tree structure. The analysis showed that the value of *warm directiveness* was a key factor differentiating the behaviour of children in preschool. The first split point, identified at a level greater than 95, revealed a group of 136

individuals. In this group, the majority—84 individuals—were parents of children who were assessed by teachers as "well-behaved", while the remaining 52 were parents of "difficult" children. This group was marked as *ID = 4* in the tree structure.

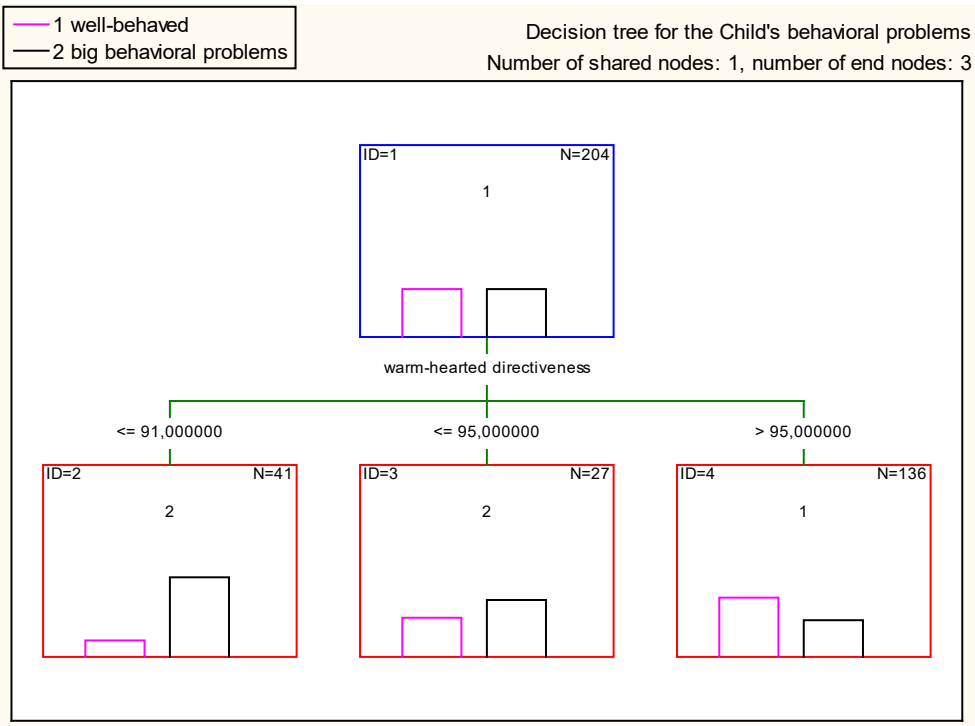


Figure 8.2. Classification tree with quantitative predictors constructed using the CHAID algorithm

The next split point concerned the value of *warm directiveness* less than or equal to 95. This category included 27 individuals. The majority—16 individuals—were parents of children who were classified as "difficult", while 11 were parents of "well-behaved" children. This group was marked as *ID = 3*.

The third split point considered a value of *warm directiveness* less than or equal to 91. This category included 41 individuals. The vast majority—34 individuals—were parents of children who exhibited "difficult" behaviour in preschool. The remaining 7 individuals were parents of "well-behaved" children. This group was marked as *ID = 2*.

The results of the analysis indicate a clear relationship between the level of parents' *warm directiveness* and the teacher's opinion of the child's behaviour. The higher the level of warm directiveness, the greater the likelihood that the child would be assessed as "well-behaved". The analysis also highlights the challenge of accurately classifying "difficult" children, as reflected in the lower accuracy for this category.

Table 8.4. Structure of the classification tree with quantitative predictors constructed using the CHAID algorithm

Node	No. of branches	Node size	Class 1 (well-behaved)	Class 2 (difficult)	Selected class	Split variable	Criterion for child 1	Criterion for child 2	Criterion for child 3	Child node 1	Child node 2	Child node 3
1	3	204	102	102	1	warm_directiveness	x ≤ 91.0000	91.0000 < x ≤ 95.0000	x > 95.0000	2	3	4
2		41	7	34	2							
3		27	11	16	2							
4		136	84	52	1							

In conclusion, the CHAID algorithm made it possible to identify key relationships between *warm directiveness* and the teacher's opinion, which may serve as a foundation for further research on the impact of parental communication on child behaviour. These findings also suggest the potential for practical application of the algorithm in the area of diagnosis and support for child development in educational settings.

8.3. Interactive Approaches to Quantitative Predictors

The purpose of interaction-based algorithms used in building classification trees with quantitative predictors is to construct a model that explains a qualitative variable based on quantitative variables. These algorithms take into account both main effects and interaction effects—i.e., the mutual influence of two quantitative predictors on a qualitative variable. Such interactions can be described in terms of mediation and moderation effects.

A *mediation effect* occurs when variable Y (the mediator) is simultaneously explained by variable X and itself explains the qualitative variable. A *moderation effect*, on the other hand, occurs when variable Y influences the relationship between variable X and the dependent variable. Mediation and moderation are the most common forms of interaction between quantitative variables, as described by Stefan Nowak (Nowak, 2007).

In classification trees with quantitative predictors that include interaction effects, the model describes the combined influence of two predictors on the qualitative (categorical) variable. In the C&RT algorithm, these interactions always result in a split into two nodes. In contrast, the CHAID algorithm allows for more complex splits—more than two branches may emerge from a single node.

This means that in the case of the C&RT algorithm, the interaction effect is limited to a maximum of two groups per tree branch. In the CHAID algorithm, an interaction effect may lead to the identification of two or more groups within a single branch. This makes CHAID more flexible in analysing complex relationships by allowing for more detailed segmentation of data based on the intensity of interaction effects between predictors.

CHAPTER 9

Regression Trees with Qualitative Predictors: Predictive Models

Regression trees with qualitative predictors are analytical tools in which the dependent variable is numerical, while the explanatory variables (predictors) are qualitative. These models resemble the approach used in analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA), where numerical dependent variables are explained using qualitative predictors (Brzeziński & Stachowski, 1984).

The first part of this chapter discusses the basic version of the C&RT algorithm, which focuses on analysing main effects. We will present how the values of a dependent variable can be explained based on qualitative predictors by dividing the data into homogeneous subgroups. The second part of the chapter focuses on the interaction effects between different levels of the predictors, analysing their influence on the dependent variable. This will illustrate how regression trees can be applied in modeling and forecasting in situations where qualitative variables play a key role.

9.1. C&RT in Regression Modeling

A decision tree was constructed using the C&RT algorithm for the variable *teaching the child rules*, taking into account two predictors: a) the child's birth order in the family and b) the child's age. *Teaching the child rules* was a numerical (interval) variable serving as the dependent variable. In turn, birth order and age, as ordinal variables, were used as predictors. Figure 9.1 illustrates the detailed structure of the decision tree, including successive splits and the mean values of the dependent variable for each group. Table 9.1 presents the full structure of the tree along with

information on the importance of predictors. The analysis showed that the variable *birth order in the family* had the greatest influence on the model construction, while *child's age* ranked second in importance.

The C&RT algorithm generated a tree in which the first branch was based on the variable *birth order in the family*. For values corresponding to middle and youngest children (value 2 or 3), the algorithm identified a group of 75 individuals (*ID2*) with a mean value of 11.81 for *teaching rules*. In contrast, group *ID3*, which included the oldest children in the family (value 1), comprised 129 individuals with a mean of 10.279 for the same variable. The mean values for both groups are shown in nodes *ID2* and *ID3* in Figure 9.1.

Further analysis by the algorithm differentiated individuals within these subgroups. In *group ID2*, subsequent splits were made based on the variable *child's age*. For five-year-old children (second level of the variable), the algorithm identified 26 individuals with a mean of 10.692 in *teaching rules* (*ID4*). For four- and six-year-olds, the algorithm distinguished a group of 49 individuals with a higher mean of 12.40, indicating a greater intensity of teaching rules (*ID5*).

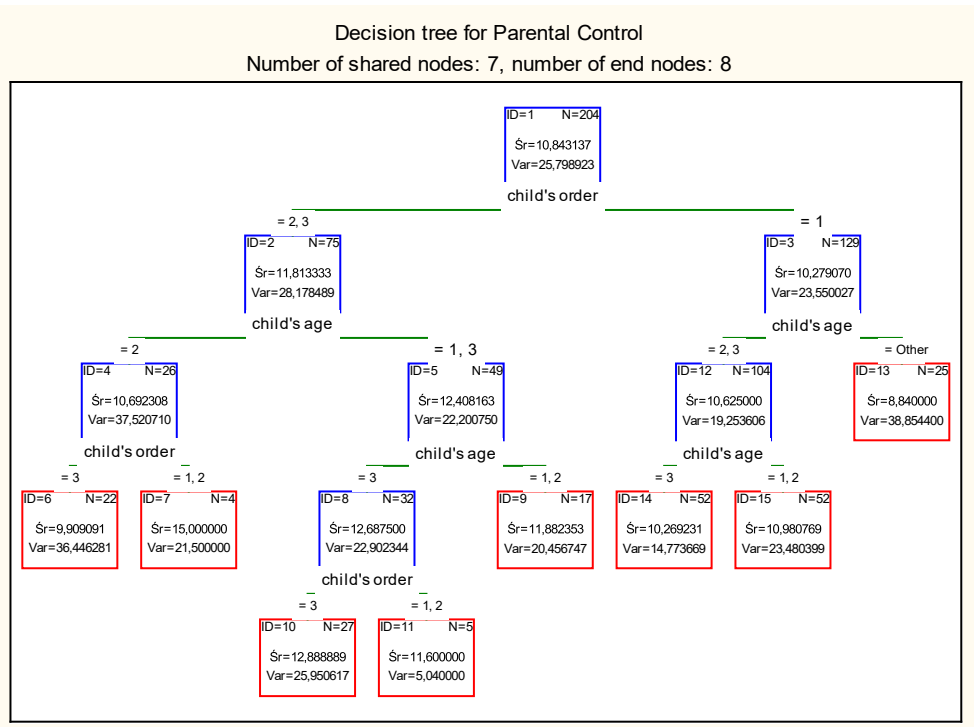


Figure 9.1. Regression tree with qualitative predictors constructed using the C&RT algorithm

In *group ID5*, the algorithm once again used the *child's age* variable for further division. Among six-year-old children (third level of the *child's age* variable), the

mean of *teaching rules* was higher (*ID8*) compared to the rest of the group (*ID9*). *Group ID9* was considered a terminal node, meaning the algorithm did not perform any further splits within it. In contrast, in *group ID8*, an additional split was made based on the variable *birth order in the family*. The analysis showed that youngest children were taught rules more intensively than older ones (*ID10* and *ID11*).

To further differentiate *group ID4*, the algorithm once again used the variable *birth order in the family*. The analysis revealed that the youngest children had the lowest mean score in being taught rules—9.909—while older children achieved a higher mean.

Group ID3, which from the beginning was characterised by a lower value in terms of *teaching rules*, was subjected to further analysis using the variable *child's age*. The algorithm showed that five- and six-year-old children were taught rules more intensively—this applied to *group ID12* (Figure 9.1). The mean value for *teaching rules* in this group was 10.625, and the algorithm classified 104 individuals into it. An additional split within this group revealed that six-year-old children were taught rules slightly less intensively than other groups, although the differences were small (*ID14* and *ID15*, Figure 9.1).

Table 9.1. Structure of the decision tree presented in Figure 9.1

Node No.	Left branch	Right branch	Node size	Node mean	Node variance	Split variable	Split constant	Class left	Class right
1	2	3	204	10.84314	25.79892	birth_order		2	3
2	4	5	75	11.81333	28.17849	child_age		2	
4	6	7	26	10.69231	37.52071	birth_order		3	
6	–	–	22	9.90909	36.44628				
7	–	–	4	15.00000	21.50000				
5	8	9	49	12.40816	22.20075	child_age		3	
8	10	11	32	12.68750	22.90234	birth_order		3	
10	–	–	27	12.88889	25.95062				
11	–	–	5	11.60000	5.04000				
9	–	–	17	11.88235	20.45675				
3	12	13	129	10.27907	23.55003	child_age		2	3
12	14	15	104	10.62500	19.25361	child_age		3	
14	–	–	52	10.26923	14.77367				
15	–	–	52	10.98077	23.48040				
13	–	–	25	8.84000	38.85440				

The decision tree model identified two distinct groups of parents: those of older children and those of younger and middle children. Each of these groups exhibits differences in their approach to teaching rules. The right side of the tree distinguished parents of the oldest children in the family, while the left side of the first branch of the algorithm focused on parents of the youngest and middle children.

The analysis performed by the algorithm indicates that, in the case of the oldest children—particularly five- and six-year-olds—parents demonstrated greater intensity in teaching rules. The older the child, the higher the level of rule-teaching, though this pattern did not apply to four-year-olds. The lowest mean score in rule-teaching was observed among four-year-old children who were the oldest among their siblings.

For the youngest and middle children in the family, rule-teaching followed different patterns. Parents taught rules to four- and six-year-olds with similar intensity. Interestingly, five-year-old children had the lowest mean level of rule-teaching. This may be related to the developmental characteristics of five-year-olds—they are highly active due to rapid motor development, while their intellectual development has not yet reached the level observed in six-year-olds. Causal thinking at this age is still emerging, which may lead parents to delay more intensive rule-teaching during this stage.

The decision tree model built using the C&RT algorithm revealed significant relationships between the child's birth order, the child's age, and the intensity of rule--teaching by parents. The key predictor proved to be *birth order in the family*, highlighting its importance in upbringing processes. The results indicate that the oldest children in the sibling group are taught rules in a more structured manner, particularly at the ages of five and six. In contrast, for the youngest and middle children, the process of rule-teaching is more varied, with noticeable differences depending on age.

A particularly interesting observation is the lower intensity of rule-teaching among five-year-old children, which may be linked to their stage of motor and cognitive development. These findings offer practical insights for parents and educators, who may adjust their approach to teaching rules depending on the child's age and position within the family.

9.2. Advanced Tree Regression Models Including Interactions

The aim of the interaction-based algorithm in constructing regression tree models is to identify combinations of levels of qualitative variables (predictors) that best explain variation in a quantitative dependent variable. These models allow for the inclusion of interaction effects between the levels of two predictors, enabling more accurate prediction of the dependent variable.

The regression trees described in this chapter illustrate the operation of interaction effects, which correspond to interaction effects in analysis of variance. In this approach, the combination of levels of two qualitative variables influences the quantitative dependent variable. Examples of such effects were discussed in detail in Chapter 7.4. The regression tree models presented here focus on explaining this type of relationship, highlighting the importance of mutual connections between qualitative variables in shaping the values of a quantitative outcome.

Introducing interactions between predictors in regression models allows for a more precise consideration of complex relationships within the data, making these techniques particularly useful in the analysis of multidimensional problems.

9.2.1. C&RT: Implementation and Case Studies

In reference to the example discussed in Chapter 7.4 and previously published, an analysis of a decision tree including interaction effects was conducted. The dependent variable was *obedience*, which was quantitative in nature, while the independent variables were *teacher's opinion* and *parent's gender*, both of which were qualitative. *Teacher's opinion* was classified into two levels: Level 1 – "well-behaved" child and Level 2 – "difficult" child. The *parent's gender* variable also had two levels: Level 1 – mother, and Level 2 – father.

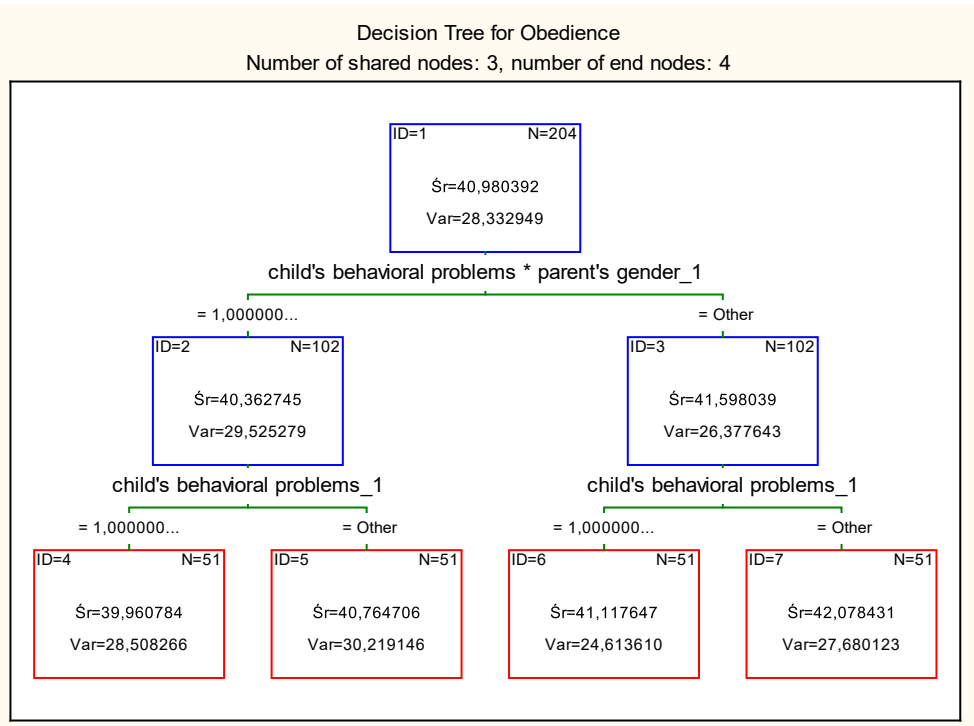


Figure 9.2. Graph of an interaction-based regression tree with qualitative predictors constructed using the C&RT algorithm

The C&RT algorithm first extracted an interaction between *teacher's opinion* and *parent's gender*, revealing a relationship between "well-behaved" children and the maternal parenting role, and between "difficult" children and the paternal parenting role. The analysis results showed that this group had a lower mean level of demanding obedience—40.36 (see *ID2*, Figure 9.2)—compared to the remaining group consisting of fathers of "well-behaved" children and mothers of "difficult" children,

whose mean was 41.59 (see *ID3*, Figure 9.2). This result is consistent with previously published findings (Szymańska, 2009).

Within the subgroup with the lower mean for demanding obedience (*ID2*), the algorithm identified another dependency. In the case of *mothers of well-behaved children* (*ID4*), the mean was lower than in the group of *fathers of difficult children* (*ID5*). Similarly, in the second part of the group (*ID3*), the algorithm indicated that *fathers of well-behaved children* had a lower mean for demanding obedience (*ID6*) than *mothers of difficult children* (*ID7*).

The strongest influence on the construction of the tree came from *teacher's opinion* at Level 1, that is, "well-behaved children". Second in importance were the interaction effects between *teacher's opinion* and *parent's gender*—specifically, the relationship between "well-behaved children" and mothers. Only in third place did the mother's gender alone appear as a relevant variable in tree construction. The structure of the tree is presented in detail in Table 9.2**.**

Table 9.2. Structure of the tree presented in Figure 9.2

Node No.	Left branch	Right branch	Node size	Node mean	Node variance	Split variable	Split constant	Split class
1	2	3	204	40.98039	28.33295	teacher_opinion* parent_gender_1		1.000000
2	4	5	102	40.36275	29.52528	teacher_opinion_1		1.000000
4	–	–	51	39.96078	28.50827			
5	–	–	51	40.76471	30.21915			
3	6	7	102	41.59804	26.37764	teacher_opinion_1		1.000000
6	–	–	51	41.11765	24.61361			
7	–	–	51	42.07843	27.68012			

The analysis conducted using the C&RT algorithm provided significant insights into the influence of *teacher's opinion* and *parent's gender* on children's level of obedience, revealing interaction effects between these variables. The results showed that the most important predictor in tree construction was *teacher's opinion*, and the key element differentiating the groups was the low level of demanding obedience among *mothers of well-behaved children*.

The decision tree enabled the identification of subgroups in which *mothers of well-behaved children* displayed particularly low mean scores, suggesting unique interaction mechanisms within this group. These results are consistent with earlier research (Szymańska, 2009), while also contributing new insights that were not apparent in the analysis of variance.

The presented tree structure (Table 9.2) and its visualisation (Figure 9.2) reveal detailed relationships and a hierarchy of variables, allowing for a better understanding of the complexity of how *teacher's opinion* and *parent's gender* influence children's behaviour. These findings provide a valuable complement to previous studies and may be applied in further predictive analyses and educational practice.

CHAPTER 10

Regression Trees with Quantitative Predictors: Methods and Applications

Regression trees with quantitative predictors are models in which both the dependent variable and the predictors (explanatory variables) are expressed in numerical form. These models offer a flexible alternative to classical regression models such as simple and multiple regression, enabling more detailed modeling of relationships between predictors and the dependent variable. Unlike traditional methods, regression trees allow for a hierarchical representation of predictor influence and the identification of interactions between them.

This chapter discusses two main approaches to the analysis of regression trees with quantitative predictors. The first part focuses on algorithms that consider only the main effects of predictors, analysing their individual impact on the dependent variable. The second part presents models that incorporate interaction effects between predictors, allowing for the identification of more complex dependencies. Each of these approaches will be illustrated with practical examples of applications.

10.1. C&RT: Analysis and Interpretation

A regression tree model for the variable *aggressive directiveness* was constructed using the C&RT algorithm. The dependent variable was quantitative, while the predictors—(a) *discrepancy*, (b) *experienced parental difficulty*, (c) *stress coping through withdrawal*, (d) *stress coping through applying pressure*, and (e) *negative representation of the child in the parent's mind*—were all numerical variables. The analysis was conducted on a sample of 107 parents of preschool-aged children.

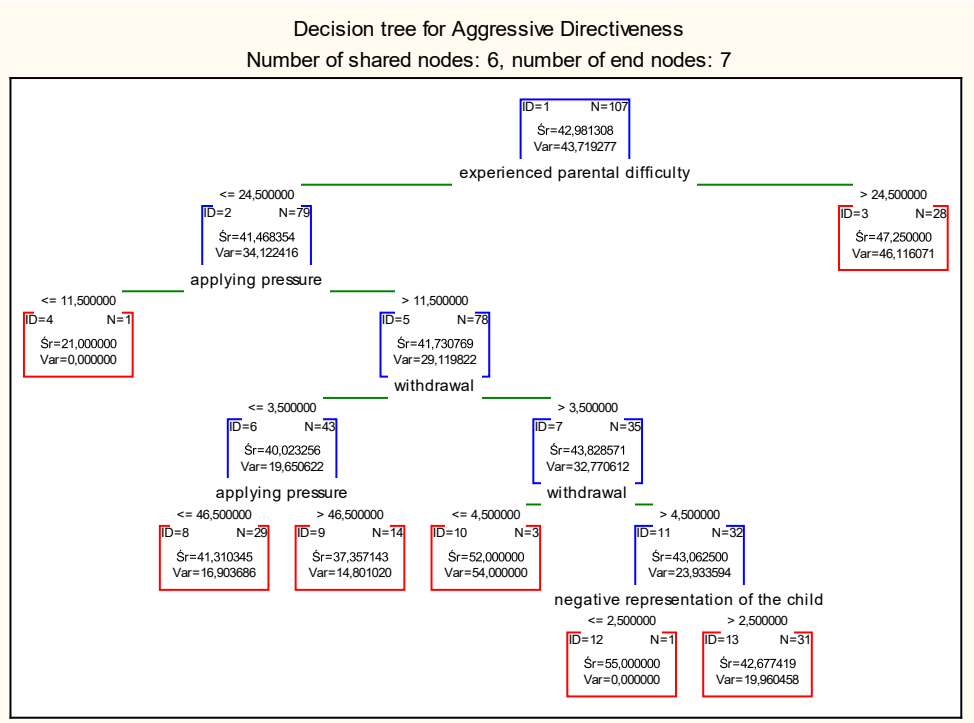


Figure 10.1. Graph of the regression tree with quantitative predictors constructed using the C&RT algorithm

The algorithm performed the first split function based on the variable *experienced parental difficulty*. Values greater than 24.5 identified *group ID=3*, consisting of 28 individuals with the highest mean score in *aggressive directiveness*, amounting to 47.25. As a homogeneous group, it was not subjected to further splits, since no additional partitioning was necessary. The analysis clearly demonstrated a relationship between high parental difficulty and an increase in aggressive communication with the child.

On the other side of the split was *group ID=2*, in which the level of *experienced parental difficulty* was less than or equal to 24.5. This group included 79 individuals and had a mean *aggressive directiveness* score of 41.46. Unlike group *ID=3*, this group underwent further divisions to allow for more detailed analysis.

The algorithm performed the next split using the variable *stress coping through applying pressure*. Values less than or equal to 11.5 identified *group ID=4*, consisting of a single individual whose *aggressive directiveness* score was 21. The remaining portion of the sample was assigned to *group ID=5*, where the mean score of *aggressive directiveness* was 41.73.

For *group ID=5*, the algorithm used the variable *stress coping through withdrawal*. Values of this variable less than or equal to 3.5 formed *group ID=6*, characterised

by a lower mean score in *aggressive directiveness*. Conversely, values greater than 3.5 placed individuals in *group ID=7*, where the mean score for *aggressive directiveness* was higher.

Group ID=6, which had a lower mean in *aggressive directiveness*, underwent a final split using the variable *applying pressure*. Values of this variable greater than 46.5 identified *group ID=9*, which had the lowest mean level of aggressive directiveness among all respondents. In contrast, values less than or equal to 46.5 formed *group ID=8*, where the mean aggressive directiveness was higher.

The group characterised by a higher mean in *aggressive directiveness* was further split using the variable *withdrawing from the educational situation*. Values less than or equal to 4.5 assigned individuals to *group ID=10*, while values greater than 4.5 formed *group ID=11*, in which the mean aggressive directiveness was higher.

Group ID=11 was ultimately split based on the variable *negative representation of the child*. Values greater than 2.5 identified *group ID=13*, which had a lower mean in *aggressive directiveness*. On the other hand, values less than or equal to 2.5 assigned individuals to *group ID=12*, where the mean aggressive directiveness was higher.

Based on the analysis, the algorithm identified several important rules. First, an increase in *parental difficulty* is associated with a higher level of *aggressive directiveness*. Second, *withdrawing from the parenting process* as a stress-coping strategy is linked to an increase in aggressive directiveness. The group of individuals with higher withdrawal scores (*ID=7*) and their successors demonstrated higher levels of *aggressive directiveness* compared to the group with lower withdrawal scores (*ID=6*) and their successors.

Table 10.1. Structure of the tree presented in Figure 10.1

Node No.	Left branch	Right branch	Node size	Node mean	Node variance	Split variable	Split constant
1	2	3	107	42.98131	43.71928	parental_difficulty	24.50000
2	4	5	79	41.46835	34.12242	stress_pressure	11.50000
4	–	–	1	21.00000	0.00000		
5	6	7	78	41.73077	29.11982	stress_withdrawal	3.50000
6	8	9	43	40.02326	19.65062	stress_pressure	46.50000
8	–	–	29	41.31034	16.90369		
9	–	–	14	37.35714	14.80102		
7	10	11	35	43.82857	32.77061	stress_withdrawal	4.50000
10	–	–	3	52.00000	54.00000		
11	12	13	32	43.06250	23.93359	negative_representation	2.50000
12	–	–	1	55.00000	0.00000		
13	–	–	31	42.67742	19.96046		
3	–	–	28	47.25000	46.11607		

Table 10.1 presents the structure of the regression tree. The most important predictor in tree construction turned out to be *discrepancy*, followed by: *stress coping through withdrawal*, *experienced parental difficulty*, *stress coping through applying pressure*, and *negative representation of the child in the parent's mind*.

The variable *discrepancy* received a high score in the estimation of predictor importance, even though it was not used in the final tree structure. This may indicate that it played a significant role in alternative splits considered by the algorithm but was ultimately excluded during the pruning process. It is important to note that the assessment of variable importance also includes those splits that were considered but not retained in the final tree.

CHAPTER 11

Statistical Significance of Nodes in Decision Trees

Although the procedure for computing decision trees does not offer a direct means to verify whether there are statistically significant differences between the groups identified in a given node or leaf, such an analysis is possible through additional steps. The tree classifies each examined individual into one of the subgroups it distinguishes and assigns them a value corresponding to that subgroup. To conduct a statistical analysis, the column containing the classification can be copied and inserted into the original dataset. For instance, in the model presented in Figure 7.3, the algorithm assigns each individual a value corresponding to their membership in a specific leaf (e.g., ID = 2, ID = 4, ID = 5). To compare whether there are statistically significant differences between group ID = 2 and group ID = 3 (formed by merging ID = 4 and ID = 5), it suffices to combine individuals belonging to ID = 4 and ID = 5 into a single group labelled ID = 3. Group ID = 3 was originally split into subgroups (ID = 4 and ID = 5) during the tree-building process. Therefore, to reconstruct the complete group ID = 3, one must merge elements ID = 4 and ID = 5 into a single group.

Subsequently, an appropriate statistical test can be conducted to determine whether statistically significant differences exist between groups ID = 2 and ID = 3. In the case of the data presented in Figure 7.3, the Mann–Whitney U test may be appropriate, as it enables the comparison of two independent groups even when the data do not meet the assumption of normal distribution. It is important to consider several key methodological aspects. Groups may only be merged if they originate from the same branch of the tree. For instance, ID = 4 and ID = 5 can be merged into ID = 3 only if they were originally part of the same group that was later

divided during tree construction. It should also be noted that statistical tests such as the Mann–Whitney U test require a minimum group size to ensure adequate statistical power. When groups are very small, the results may be unreliable, and the test may fail to detect actual differences.

If the analysis involves more than two groups, the Kruskal–Wallis test can be applied, which allows for the identification of differences between several groups without assuming a normal distribution.

Let us assume that groups ID = 2 and ID = 3 contain 50 and 57 cases respectively, and that the variable values in each group are not normally distributed. In such a case, the Mann–Whitney U test allows for comparing the mean ranks of both groups. The test result will indicate whether the differences between the groups are statistically significant, which may provide additional insight into the structure of the decision tree.

In summary, the statistical significance of nodes in decision trees can be analysed through additional steps of statistical testing. Such an approach enables a deeper understanding of the differences between groups and their statistical relevance, while simultaneously considering the complexity of the relationships between predictors. However, it requires caution in selecting analytical methods and interpreting results within the context of the overall tree structure. With proper application of these methods, it is possible to obtain more precise and valuable conclusions from the analysis.

CHAPTER 12

Data Imputation Using Decision Trees

Decision trees can be successfully employed to fill in missing data within datasets. As they are capable of predicting group membership or variable values based on other variables, these models—especially those demonstrating high goodness of fit—constitute an effective tool for data imputation. Decision trees, as models of mathematical induction belonging to artificial intelligence algorithms, generate rules that enable the prediction of dependent variable values based on predictors. Depending on the type of tree, they can be used either to predict the values of quantitative variables or to classify cases into groups in the case of qualitative variables.

For example, if a dataset lacks values for the variable "age", decision trees can be used to predict the missing values based on other variables such as level of education, place of residence, or income. In the case of qualitative variables such as "gender", decision trees can predict missing values using occupational preferences, interests, or other descriptive variables. Thus, decision trees constitute a versatile tool that enables data completion in various research contexts.

However, when applying this procedure, special caution must be exercised to avoid the accumulation of prediction errors. In particular, one should refrain from reusing variables that were previously predicted in order to fill in missing data. Such a situation could lead to a vicious circle in which imputed values begin to explain other variables, thereby distorting the results of the analysis. It is therefore essential to thoroughly analyse the dataset structure and logically exclude the reuse of the same variables.

It should also be emphasised that only well-fitted models with high accuracy should be used for data imputation. This can be assessed using a designated test dataset. Verifying model performance on such a test set helps to avoid overfitting,

which could otherwise lead to inadequate predictions of missing data. These tests are particularly important when data imputation affects subsequent statistical analyses or predictive models.

Alternative imputation methods—such as filling in missing values with means, medians, kNN algorithms, or regression—are also used; however, decision trees offer an advantage when dealing with data that exhibit complex relationships between variables. For instance, whereas imputation with the mean does not take into account variable context, decision trees utilise intricate interactions between predictors, allowing for more precise imputations.

In cases where a dataset contains a large number of missing values, it is also necessary to consider the minimum group size in order to ensure adequate statistical power. For example, if the group size is too small, the prediction results may be unreliable, and the imputed values may fail to reflect the true distribution of the variable in the population.

In summary, decision trees represent an effective tool for completing missing data in both quantitative and qualitative variables. Their application enables the use of complex relationships between variables to more accurately impute missing values. Nevertheless, this procedure requires careful selection of variables and validation of results to avoid potential errors and to obtain reliable outcomes.

PART III

Applications of Association and Clustering Algorithms and Link Analysis in Data Analysis

CHAPTER 13

Fundamentals of Association Algorithms

Association algorithms, also referred to as association rule mining algorithms, are among the key techniques in the field of data mining and artificial intelligence. Their primary aim is to identify recurring patterns, dependencies, or relationships within large datasets. These algorithms are most commonly applied in the analysis of purchase transactions, where they allow researchers to discover which products are frequently bought together (Srikant & Agrawal, 1995).

Association algorithms are used to uncover relationships between various elements in datasets. A typical example is market basket analysis, in which the objective is to identify sets of products that are frequently purchased together by customers in retail settings. This information is subsequently used to optimise store layouts, create promotional offers, or generate product recommendations (Srikant & Agrawal, 1995).

Association algorithms are an integral part of artificial intelligence and data mining. While they are not classified as AI algorithms per se, they constitute a tool that supports the development of intelligent recommendation systems. These systems analyse transactional data to predict user preferences and deliver personalised recommendations.

The most popular association algorithms include the Apriori algorithm, the FP-Growth algorithm, and the Eclat algorithm. Apriori is one of the earliest and best-known association algorithms. It is based on the principle that every subset of a frequent itemset must also be frequent. This algorithm generates candidate frequent itemsets and subsequently verifies their frequency within the dataset (Srinadh, 2022). The FP-Growth (Frequent Pattern Growth) algorithm is an alternative to Apriori that eliminates the need to generate a large number of candidates. It uses a tree

structure known as an FP-Tree to represent frequent itemsets (Srinadh, 2022). The Eclat algorithm (Equivalence Class Clustering and bottom-up Lattice Traversal) is another efficient association algorithm that operates by breaking the problem down into smaller subproblems and processing them concurrently (Srinadh, 2022).

Association algorithms differ from classification algorithms. Whereas classification algorithms are designed to assign data to predefined classes, association algorithms focus on uncovering relationships among elements in the dataset without prior knowledge of these relationships. Classification involves predicting specific labels, whereas association algorithms aim to discover patterns that may not be immediately apparent.

Association algorithms are widely used in various domains, including marketing and retail, finance, medicine, and computer science. In marketing and retail, they are used for market basket analysis, product recommendations, and store layout optimisation (Karthiyayini & Balasubramanian, 2016). In finance, they support the detection of fraudulent transactions and the analysis of investment portfolios (Sarno et al., 2015). In medicine, they aid in analysing correlations between medications and side effects, as well as supporting medical diagnoses (Chimieski & Fagundes, 2013). In computer science, they are applied to database optimisation and recommendation systems in content management platforms (Kudriavtsev et al., 2023).

Association algorithms are a powerful tool in data mining, enabling the discovery of hidden patterns and relationships. In the following chapters, selected association algorithms will be discussed in greater detail, along with their applications and implementation.

13.1. Basket Analysis: Market Basket Analysis (MBA) Method

Market Basket Analysis (MBA) is a method used for analyzing transactional data to discover associative patterns between products (Karthiyayini & Balasubramanian, 2016). In the context of psychology and consumer behavior, MBA helps to understand which products are purchased together, potentially revealing customer preferences and shopping habits. The most commonly used algorithms for conducting this analysis include Apriori, FP-Growth, and Eclat, which effectively identify patterns and relationships within large datasets. Let us take a closer look at the assumptions and formal representations of this method, and explore how it can be applied in practice.

The information presented in this chapter is based on publicly available knowledge in the field of market basket analysis, particularly the works of authors such as Srinadh (2022), Karthiyayini and Balasubramanian (2016), Kudriavtsev et al. (2023), Alawadh and Barnawi (2022), and Hoque et al. (2024).

Market Basket Analysis relies on transactional data, which constitutes a collection of all transactions made by customers. Each transaction T_i is a set of items

from the universe of items $I = \{i_1, i_2, ..., i_n\}$. In other words, each transaction contains various combinations of products available in the store's offering. The complete set of transactions can be denoted as $D = \{T_1, T_2, ..., T_n\}$, where $T_n \subseteq I$.

Transactional data is often represented in the form of a binary matrix. In this matrix, each row corresponds to one transaction, and each column corresponds to a specific product. A value of 1 in a cell indicates that a given product was present in the transaction, while a value of 0 indicates its absence.

Market Basket Analysis employs several key measures to assess the relationships between products. The most important of these include *support*, *confidence*, and *lift* (Alawadh & Barnawi, 2022; Hoque et al., 2024).

Support for an itemset $X \subseteq I$ is defined as the proportion of transactions in D that contain X. Mathematically, it can be expressed as:

$$support(X) = \frac{|\{T_i \in D : X \subseteq T_i\}|}{|D|}$$

This means that *support* measures how frequently a given set of products appears within the entire transaction dataset. For example, if 30% of all transactions include bread, then the support for bread is 0.3.

Confidence for the association rule $X \Rightarrow Y$ (where X, $Y \subseteq I$ and $X \cap Y = \emptyset$) is defined as the proportion of transactions that contain X and also contain Y. Mathematically, it is expressed as:

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

Confidence indicates the conditional probability that product Y will be purchased given that product X has already been purchased. For example, if 50% of transactions containing milk also include bread, then the confidence for the rule {milk} ⇒ {bread} is 0.5.

Lift for the association rule $X \Rightarrow Y$ measures the strength of the rule by adjusting for the frequency of occurrence of Y. It is the ratio of *confidence* to the *support* of product Y. Mathematically, it is expressed as:

$$lift(X \Rightarrow Y) = \frac{confidence(X \Rightarrow Y)}{support(Y)} = \frac{support(X \cup Y)}{support(X) \times support(Y)}$$

Lift allows us to assess whether the presence of product X increases (or decreases) the likelihood of purchasing product Y compared to a scenario in which the products are statistically independent. For example, if the lift is 1.67, this means that the presence of product X increases the likelihood of purchasing product Y by 67% above the expected baseline. Below is a detailed explanation of different cases for various lift values:

1. **Lift = 1:**
   Meaning: Products *X* and *Y* are statistically independent.
   Interpretation: The presence of product *X* has no effect on the likelihood of purchasing product *Y*. The probability of purchasing *Y* in the presence of *X* is the same as in its absence.
   Example: If the lift for the rule {milk} ⇒ {bread} is 1, it means that customers buy bread with the same frequency regardless of whether they purchased milk.

2. **Lift > 1:**
   Meaning: Products *X* and *Y* are positively correlated.
   Interpretation: The presence of product *X* increases the likelihood of purchasing product *Y*. The association rule is significant and indicates a relationship between the products.
   Example: If the lift for the rule {milk} ⇒ {bread} is 1.5, it means that purchasing milk increases the likelihood of purchasing bread by 50% above the expected level.

3. **Lift < 1:**
   Meaning: Products *X* and *Y* are negatively correlated.
   Interpretation: The presence of product *X* decreases the likelihood of purchasing product *Y*. The products may be perceived as substitutes or otherwise unfavorably related.
   Example: If the lift for the rule {milk} ⇒ {bread} is 0.8, it means that purchasing milk decreases the likelihood of purchasing bread by 20% compared to the expected frequency of bread purchases.

Table 13.1 presents the interpretation of lift values.

Table 13.1. Interpretation of Lift Values

Lift Value	Meaning	Interpretation
Lift = 1	Products are independent	Purchasing X does not affect the likelihood of purchasing Y.
Lift > 1	Products are positively correlated	The presence of X increases the likelihood of purchasing Y.
Lift < 1	Products are negatively correlated	The presence of X decreases the likelihood of purchasing Y.

Let us assume the following transactional dataset:

$$D = \{\{A, B\}, \{B, C, D\}, \{A, C\}, \{A, B, C\}, \{B, C\}\}$$

To understand how the individual measures work, let us analyze this dataset.

Support for item *A*:

$$support(A) = \frac{|\{\{A, B\},\{A, C\}, \{A, B, C\}\}|}{|D|} = \frac{3}{5} = 0{,}6$$

This means that 60% of the transactions include item *A*.

Confidence for the rule $A \Rightarrow C$:

$$confidence(A \Rightarrow C) = \frac{support(A \cup C)}{support(A)} = \frac{\frac{2}{5}}{\frac{3}{5}} = 0{,}6$$

This means that in 67% of the cases when item *A* is purchased, item *C* is also purchased.

Lift for the rule $A \Rightarrow C$:

$$lift(A \Rightarrow C) = \frac{confidence(A \Rightarrow C)}{support(C)} = \frac{0{,}67}{0{,}8} = 0{,}84$$

A lift below 1 suggests that the presence of item *A* decreases the likelihood of purchasing item *C* below its expected frequency.

In the context of Market Basket Analysis, graphs are frequently used to visualize relationships between products. In such graphs, **nodes** represent individual products, and **edges** between them represent associative dependencies (STATISTICA Electronic Manual, 2012).

1. **Antecedent:** The antecedent is the product or set of products that appear on the left-hand side of an association rule. In the rule $X \Rightarrow Y$, *X* is the antecedent. In the graph, the antecedent is a node from which a directed edge leads to the consequent.
2. **Consequent:** The consequent is the product or set of products on the right-hand side of an association rule. In the rule $X \Rightarrow Y$, *Y* is the consequent. In the graph, the consequent is a node to which an edge from the antecedent points.
3. **Edges:** The edges in the graph represent association rules. They are often directed, indicating the direction of the association (from antecedent to consequent). Frequently, the edges are also labeled with the values of *confidence* or *lift*, to visualize the strength and quality of the association.

Let us assume we are analyzing association rules for products *A*, *B*, and *C* based on the dataset above. We may derive the following rules:

- **A** ⇒ **C** with confidence 0.67 and lift 0.84
- **B** ⇒ **C** with confidence 0.8 and lift 1.0
- **A** ⇒ **B** with confidence 0.67 and lift 0.83

In a graph, these rules could be represented as follows:

- Node **A** with directed edges to **C** and **B**
- Node **B** with a directed edge to **C**
- Edges labeled with *confidence* and *lift* values

Let us summarize: *support*, *confidence*, and *lift* are key indicators in Market Basket Analysis that help to understand dependencies between products. *Support* measures how frequently a given set of products appears in the transaction dataset. *Confidence* assesses the probability of purchasing one product when another product has already been purchased. *Lift*, in turn, measures the strength of the association, correcting for the effect of overall product popularity.

Despite similarities, *lift* and *correlation* differ in several key aspects:

1. **Scope of application**: Correlation is universal and can be used in any statistical analysis where the relationship between quantitative variables is of interest. In contrast, *lift* is a specialized measure used primarily in basket analysis, where the data is transactional and typically discrete.
2. **Type of data**: Correlation requires numerical data that can be compared linearly, whereas *lift* operates on binary data (presence or absence of a product in a transaction).
3. **Interpretation**: Correlation indicates the strength and direction of a linear relationship (e.g., positive correlation means that an increase in one variable is associated with an increase in another), whereas *lift* shows whether the frequency of co-occurrence of two items exceeds the expected frequency under the assumption of independence.
4. **Scalability**: *Lift* is more intuitive in the context of basket analysis, because it directly accounts for the frequency of individual items in the transaction set. Correlation can be difficult to interpret with such data, because it does not take into account the specifics of transactions.

For example, in the analysis of store transactions, *lift* allows one to assess whether products X and Y are purchased together more often than would be expected based on their individual popularity. In contrast, *correlation* does not take this specificity into account and may lead to misleading conclusions when applied to low-frequency data.

In summary, the choice between *lift* and *correlation* depends on the type of data and the purpose of the analysis. *Lift* is more precise in basket analysis, whereas *correlation* works better in general statistical analyses, especially with numerical data.

Market Basket Analysis (MBA) is a highly flexible analytical tool that can be applied to categorical variables with different numbers of levels. Nevertheless, there are certain limitations and challenges that must be considered when using it.

First and foremost, the effectiveness of basket analysis increases with the size of the dataset. The greater the number of transactions or observations, the more reliable and useful the results of the analysis become. In smaller datasets, the identified associations may be random or insufficiently representative, which leads to potential errors in interpreting the results.

Furthermore, although categorical variables may have any number of levels, an increase in the number of these levels leads to a significant increase in the complexity of the analysis. In practice, variables with a large number of levels may cause a data sparsity problem, where individual levels are rarely observed, making it difficult to identify statistically significant associations. Additionally, an increase in the number

of variable levels results in a combinatorial explosion of combinations that must be analyzed, which significantly increases the demand for computational resources and processing time.

Another aspect is the interpretation of the results of basket analysis. An increase in the number of levels of categorical variables complicates the interpretation of the results, which may require advanced analytical methods and specialized knowledge. Focusing on variables with a smaller number of levels can facilitate understanding and the use of the analysis results, reducing the risk of overinterpretation or erroneous conclusions.

Finally, it is crucial to assess the practical significance of the identified associations. Even with a large number of variable levels, not all discovered associations will be relevant from a research or practical standpoint. It is necessary to conduct additional verification analyses to confirm the significance and usefulness of the results.

13.1.1. Examples: Association Analysis in the Context of Parenting and Children's Behavior Using Market Basket Analysis (MBA)

Market Basket Analysis can be applied not only to the analysis of purchase data, but also to measuring the level of association between psychological variables. This method is particularly useful when we seek to understand relationships and patterns in data that are not quantitative in nature but qualitative. For instance, in psychology, basket analysis can be used to examine relationships between various personality traits, psychological diagnoses, behaviors, or preferences.

Qualitative psychological variables are those that are not expressed in numerical form but in the form of categories or levels. These may include, for example, different personality types (e.g., introverted, extroverted), emotional states (e.g., happy, sad), coping styles (e.g., avoidance, confrontation), or behavioral classifications (e.g., aggressive, passive).

Basket analysis in psychology can assist in identifying patterns of co-occurrence of such variables. For example, we may discover that a certain personality type often co-occurs with a specific coping style, or that particular emotional states are frequently associated with certain behaviors. This type of analysis can provide valuable insights for therapists, psychologists, and researchers, helping to better understand complex patterns of behavior and interpersonal relationships.

An example of the application of basket analysis in a psychological context may be the study of the relationship between a teacher's opinion and the child's gender. In this case, the teacher's opinion could have levels such as "well-behaved child" and "difficult child", while the child's gender could be a two-level variable ("girl" and "boy"). This approach is relatively simple, and the number of variable levels is limited, which allows for an effective implementation of basket analysis.

However, adding more levels to the variables—such as different categories of behavior (e.g., "well-behaved", "moderate", "difficult", "very difficult") and additional

characteristics of children (e.g., age, academic performance)—leads to a significant increase in the number of combinations that must be analyzed. Such complexity requires advanced data processing techniques and substantially greater computational resources.

Therefore, although basket analysis is an extremely versatile analytical tool that can be applied to categorical variables with varying numbers of levels, certain practical limitations exist. These limitations relate to dataset size, data sparsity, computational complexity, and difficulties in interpreting the results. For this reason, it is essential to maintain an appropriate scale of analysis and ensure data representativeness, which is necessary for obtaining reliable and useful results.

First Example: Association Analysis Between Teacher's Opinion and Child's Gender Using Market Basket Analysis (MBA)

In the example discussed, associations were built between the two-level variable *teacher's opinion*, where the first level denotes "well-behaved child" and the second level "difficult child", and the child's gender, where the first level denotes "girl" and the second "boy".

Table 13.2 presents the results of the basket analysis conducted for the variables "teacher's opinion" and "child's gender". The analysis was carried out with a minimum support of 20%, a minimum confidence of 50%, and a minimum correlation of 50%. The maximum length of antecedent and consequent was set to 10. Table 13.2 provides a summary of association rules from the basket analysis for the variables *teacher's opinion – child's gender*.

Table 13.2. Summary of Association Rules in the Basket Analysis for the Variables

Educator's Opinion – Child's Gender
Summary of association rules
Min. support = 20.0%; Min. confidence = 50.0%; Min. correlation = 50.0%
Max. antecedent length = 10, Max. consequent length = 10

Antecedent	⇒	Consequent	Support (%)	Confidence (%)	Correlation (%)
1	Educator's opinion == obedient child	⇒	Child's gender == girl	30.39216	60.78431
2	Educator's opinion == difficult child	⇒	Child's gender == boy	32.35294	64.70588
3	Child's gender == girl	⇒	Educator's opinion == well-behaved child	30.39216	63.91753
4	Child's gender == boy	⇒	Educator's opinion == difficult child	32.35294	62.26415

The first association rule, in which the educator's opinion indicates an "obedient child" and implies "girl", has a support of 30.39%. This means that 30.39% of all observations contain this combination. A confidence of 60.78% indicates that

in 60.78% of cases where the educator evaluates the child as obedient, the child is a girl. The correlation of 62.33% (a regular *r* correlation, presented unusually in percentages rather than as an R^2 value) suggests a moderate association between being a girl and being evaluated as an obedient child by the educator.

The second association rule, where the educator's opinion indicates a "difficult child" and implies "boy", shows a support of 32.35%. This means that 32.35% of all observations contain this combination. A confidence of 64.71% means that in 64.71% of cases where the educator evaluates the child as difficult, the child is a boy. A correlation of 63.47% indicates a moderate association between being a boy and being evaluated as a difficult child by the educator.

The third association rule, where the child's gender indicates "girl" and implies the educator's opinion of "obedient child", has a support of 30.39%, meaning that 30.39% of all observations contain this combination. A confidence of 63.92% indicates that in 63.92% of cases where the child is a girl, the educator evaluates her as obedient. The correlation of 62.33% suggests a moderate association between being a girl and being evaluated as an obedient child by the educator.

The fourth association rule, where the child's gender indicates "boy" and implies the educator's opinion of "difficult child", has a support of 32.35%. This means that 32.35% of all observations contain this combination. A confidence of 62.26% indicates that in 62.26% of cases where the child is a boy, the educator evaluates him as difficult. The correlation of 63.47% indicates a moderate association between being a boy and being evaluated as a difficult child by the educator.

In summary, the basket analysis clearly reveals significant associative patterns between the child's gender and the educator's opinion. Educators most frequently associate girls with obedient behaviour and boys with difficult behaviour. These patterns may have important implications for education and intervention strategies in preschool institutions, suggesting the need for further research and actions aimed at minimising potential biases and stereotypes related to the child's gender.

Table 13.3. Frequency of Itemsets for the Variables Educator's Opinion – Child's Gender

Frequencies of itemsets calculated
Min. support = 20.0%; Min. confidence = 50.0%; Min. correlation = 50.0%
Max. antecedent length = 10, Max. consequent length = 10

Popular Itemsets	Frequency	Support (%)
1	Educator's opinion == obedient child	102.0000
2	Educator's opinion == difficult child	102.0000
3	Child's gender == girl	97.0000
4	Child's gender == boy	106.0000
5	Educator's opinion == obedient child, child's gender == girl	62.0000
6	Educator's opinion == difficult child, child's gender == boy	66.0000

Table 13.3 presents the frequency of itemsets for the variables "educator's opinion" and "child's gender". These results were computed with a minimum support of 20%, minimum confidence of 50%, and minimum correlation of 50%. The maximum length for both antecedent and consequent was set to 10.

The first itemset, *educator's opinion = obedient child*, has a frequency of 102, indicating that 102 cases in the dataset include this educator's opinion. The support of this set is 50%, meaning that half of all observations contain the opinion that the child is obedient.

The second itemset, *educator's opinion = difficult child*, also has a frequency of 102 and support of 50%. This indicates that half of all observations in the dataset contain the opinion that the child is difficult.

The third itemset, *child's gender = girl*, appears 97 times in the dataset, yielding a support of 47.55%. This means that 47.55% of all cases in the dataset concern girls.

The fourth itemset, *child's gender = boy*, has a frequency of 106, which constitutes a support of 51.96%. This means that 51.96% of all cases in the dataset concern boys.

The fifth itemset, *educator's opinion = obedient child, child's gender = girl*, has a frequency of 62, which translates to a support of 30.39%. This means that 30.39% of all cases contain both the educator's opinion that the child is obedient and the information that the child is a girl.

The sixth itemset, *educator's opinion = difficult child, child's gender = boy*, appears 66 times, resulting in a support of 32.35%. This means that 32.35% of all cases contain both the educator's opinion that the child is difficult and the information that the child is a boy.

Interpretation of these results indicates certain patterns in the data concerning educators' opinions and children's gender. It is notable that educators' opinions are evenly divided between "obedient child" and "difficult child", with each itemset appearing in 50% of the cases. Furthermore, the gender distribution suggests that boys are slightly more represented in the data than girls (52% vs. 48%). The combinations of educators' opinions and children's gender show that educators more frequently associate girls with positive assessments (obedient) and boys with negative ones (difficult), which may suggest the presence of certain stereotypes or biases in evaluating children's behaviour.

In summary, the results from both tables demonstrate strong associations between the child's gender and the educators' opinions on their behaviour. The discovered patterns may have important implications for developing educational and intervention strategies aimed at reducing gender bias and stereotypes in the evaluation of children by educators.

Figure 13.1 presents a network of association rules for the variables "educator's opinion" and "child's gender". This network visualizes the relationships between the analyzed variables, indicating both the frequency of co-occurrence of particular variable levels (support) and the strength of these relationships (confidence).

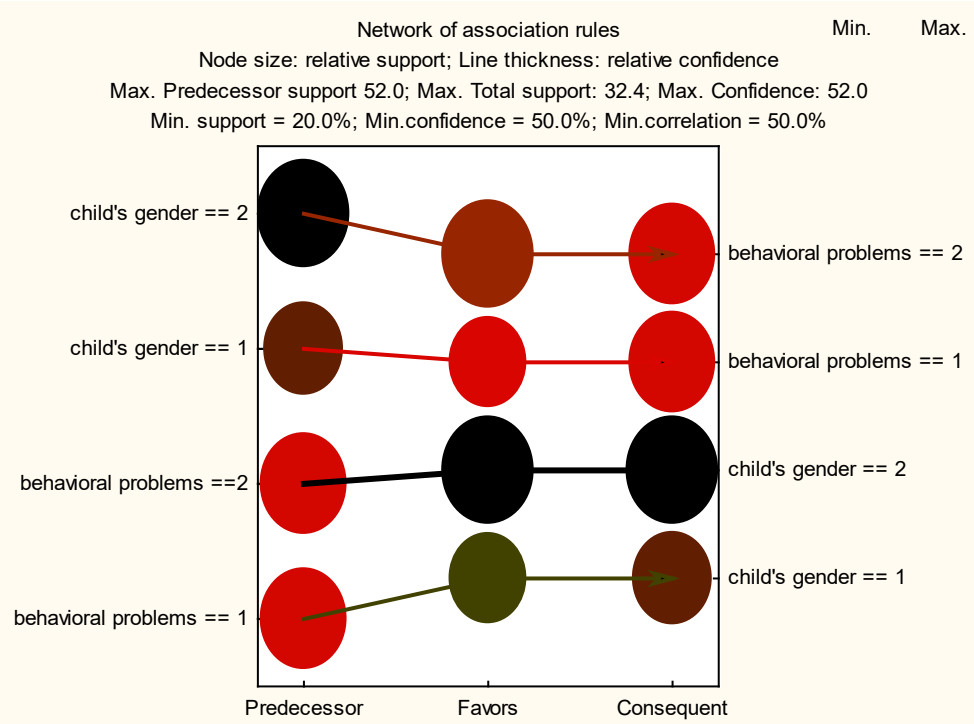


Figure 13.1. Association rule graph for the variables educator's opinion – child's gender

The nodes in the figure represent the levels of the variables "educator's opinion" and "child's gender". The variable values are labeled accordingly, e.g., "child's gender = boy", "child's gender = girl", "educator's opinion = difficult", and "educator's opinion = well-behaved". The nodes are visualized using different colors and sizes, reflecting relative support – the larger the node, the more frequent the occurrence of a given variable level in the dataset.

The edges connecting the nodes represent association rules. The thickness of these lines is proportional to the value of confidence, meaning that the thicker the line, the stronger the association between the given variable levels. For example, the thickness of the line connecting "child's gender = boy" with "educator's opinion = difficult" indicates high confidence, which means that in many cases, when the child is a boy, the educator evaluates them as difficult.

The analysis of the figure reveals several key patterns. First, the node representing "child's gender = boy" is larger than the node for "child's gender = girl". This indicates that there are more boys than girls in the analyzed dataset. At the same time, the line connecting "child's gender = boy" with "educator's opinion = difficult" is thick, indicating high confidence – boys are frequently assessed as difficult children. Similarly, the node "child's gender = girl" is connected to "educator's opinion = well-behaved" with a thick line as well, indicating that girls are frequently assessed as well-behaved children.

The figure also illustrates the structure of the network, showing that educators' opinions about children are strongly related to the child's gender. Educators tend to attribute difficult behavior to boys and well-behaved behavior to girls.

In summary, the figure clearly demonstrates the existence of strong associations between a child's gender and educators' opinions concerning their behavior. Boys are more often evaluated as difficult children, and girls as well-behaved children. These patterns suggest the presence of gender stereotypes in the assessment of children's behavior by educators. Further research could investigate the causes of these stereotypes and their impact on children's development in educational settings.

Second Example: Analysis of the Associations Between Parental Upbringing Difficulty, Child's Gender, Educator's Opinion, and Birth Order

The second sample analysis was conducted for four variables: parental upbringing difficulty, child's gender, educator's opinion, and the child's birth order within the family. *Parental upbringing difficulty* is a five-level variable, with the following levels: the first level indicates "very well-behaved child", the second level "rather well-behaved", the third level "it varies", the fourth level "I have some problems with them", and the fifth level "considerable upbringing difficulties". *Educator's opinion* is a two-level variable, with the levels defined as "well-behaved child" (level one) and "difficult child" (level two). *Child's gender* is also a two-level variable, where the first level denotes "girl" and the second level "boy". *Birth order in the family* is a three-level variable, with levels defined as: first level – "eldest child", second level – "middle child", and third level – "youngest child".

The aim of the analysis was to examine the relationships between these variables using the Market Basket Analysis (MBA) method to identify association patterns. The goal was to understand how parental perceptions of upbringing difficulty correlate with educators' opinions, the child's gender, and their birth order within the family. The results of this analysis may provide valuable insights into the impact of various factors on how children's behavior is perceived and evaluated by both parents and educators, which is crucial for developing effective educational and upbringing strategies.

A summary of the association rules is presented in Table 13.3 and visualized in Figure 13.2. For the variable *upbringing difficulty* at the level "rather well-behaved" associated with the educator's opinion that the child is "well-behaved", support was 32.35%, and confidence was 64.70%. This means that in 32.35% of cases where educators assessed a child as "well-behaved", the parents also evaluated the child as "rather well-behaved", which occurred in 64.70% of those children's cases.

Table 13.3. Summary of Association Rules from Market Basket Analysis for the Variables Educator's Opinion – Child's Gender – Birth Order – Upbringing Difficulties

Summary of association rules
Min. support = 20.0%; Min. confidence = 50.0%; Min. correlation = 50.0%
Max. antecedent size = 10; Max. consequent size = 10

Antecedent	⇒	Consequent	Support (%)	Confidence (%)	Correlation (%)
Educator's opinion == well-behaved	⇒	Child's gender == girl	30.39216	60.78431	62.33124
Educator's opinion == well-behaved	⇒	Rather well-behaved	32.35294	64.70588	64.70588
Educator's opinion == well-behaved	⇒	Eldest	32.35294	64.70588	57.53724
Educator's opinion == difficult	⇒	Child's gender == boy	32.35294	64.70588	63.47328
Educator's opinion == difficult	⇒	It varies	26.96078	53.92157	61.27017
Educator's opinion == difficult	⇒	Eldest	30.88235	61.76471	54.92191
Child's gender == girl	⇒	Educator's opinion == well-behaved	30.39216	63.91753	62.33124
Child's gender == girl	⇒	Rather well-behaved	24.50980	51.54639	50.26713
Child's gender == girl	⇒	Eldest	34.31373	72.16495	62.57737
Child's gender == boy	⇒	Educator's opinion == difficult	32.35294	62.26415	63.47328
It varies	⇒	Educator's opinion == difficult	26.96078	69.62025	61.27017
Youngest	⇒	Child's gender == boy	20.58824	63.63636	50.21395
Rather well-behaved	⇒	Educator's opinion == well-behaved	32.35294	64.70588	64.70588
Rather well-behaved	⇒	Child's gender == boy	25.49020	50.98039	50.00925
Rather well-behaved	⇒	Eldest	33.82353	67.64706	60.15256
Eldest	⇒	Educator's opinion == well-behaved	32.35294	51.16279	57.53724
Eldest	⇒	Child's gender == girl	34.31373	54.26357	62.57737
Eldest	⇒	Rather well-behaved	33.82353	53.48837	60.15256
Educator's opinion == well-behaved, Eldest	⇒	Child's gender == girl	21.56863	66.66667	54.99141
Educator's opinion == well-behaved, Eldest	⇒	Rather well-behaved	22.05882	68.18182	54.84543
Child's gender == girl, Eldest	⇒	Educator's opinion == well-behaved	21.56863	62.85714	52.07192
Rather well-behaved, Eldest	⇒	Educator's opinion == well-behaved	22.05882	65.21739	53.63989

Upbringing difficulty at the level "rather well-behaved" is also associated with the child being the eldest in the family. The support for this relationship was 33.82%, and the confidence was 67.64%. This means that in 33.82% of cases, a child assessed as "rather well-behaved" is the eldest in the family, and this relationship holds in 67.64% of those cases.

The level of upbringing difficulty described by parents as "it varies" is most strongly associated with the educator's opinion that the child is "difficult". The support for this relationship was 26.96%, and the confidence was 69.62%. This indicates that 26.96% of the children rated by parents as "it varies" are also rated as "difficult" by educators in 69.62% of those cases.

The educator's opinion that the child is "difficult" is also associated with the child being a "boy". The support for this relationship was 32.35%, and the confidence was 64.70%. This means that in 32.35% of cases, a child assessed as "difficult" is a boy, and this relationship appears in 64.70% of those cases.

Similarly, the educator's opinion that the child is "well-behaved" is associated with the child being a girl. The support for this relationship was 30.39%, and the confidence was 60.78%. This means that in 30.39% of cases, a child assessed as "well--behaved" is a girl, and this relationship appears in 60.78% of those cases.

The association between the educator's opinion that the child is "well-behaved", the child being the "eldest" in the family, and the parental evaluation of the child as "rather well-behaved" also emerged. The support was 22.05%, and the confidence was 68.18%. This indicates that in 22.05% of cases, a child assessed by the educator as "well-behaved" and being the eldest is also assessed by parents as "rather well--behaved", which occurs in 68.18% of such cases.

The dataset also revealed an association between being the eldest child and being a girl. The support for this relationship was 34.31%, and the confidence was 72.16%. This means that in 34.31% of cases, the eldest child in the family is a girl, and this relationship holds in 72.16% of those cases.

The association between the educator's opinion that the child is "well-behaved", the child being "eldest", and being a "girl" had a support of 21.56% and a confidence of 66.67%. This means that in 21.56% of cases, a child assessed as "well-behaved", being the eldest, and being a girl, confirms this relationship in 66.67% of the cases.

In summary, the market basket analysis clearly shows strong associations between parental upbringing difficulty, the child's gender, the educator's opinion, and the child's birth order within the family. These results suggest that educators' and parents' evaluations are strongly related to both gender and birth order, which may have important implications for educational and upbringing strategies.

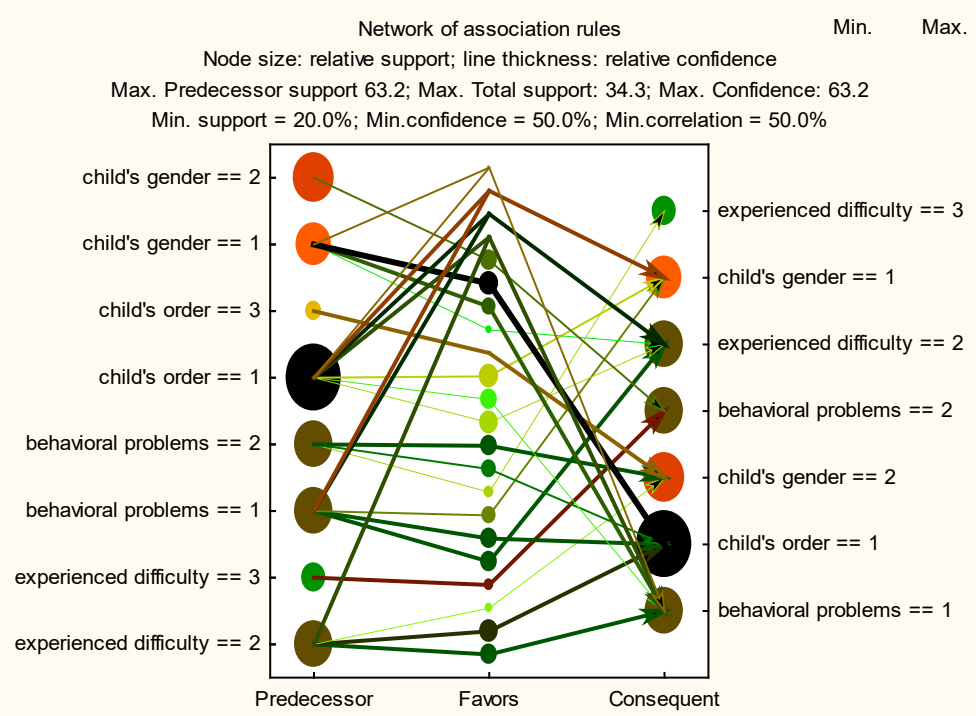


Figure 13.2. Association Rule Graph from Market Basket Analysis for the Variables Educator's Opinion – Child's Gender – Birth Order – Upbringing Difficulties

The new analysis, conducted on a broader set of variables, confirmed the rules derived from the two previously discussed variables. Additionally, it revealed some interesting associations related to the child's birth order within the family. The eldest child is associated with being "well-behaved"; however, an interaction emerges, as the eldest child in this dataset was most frequently a girl, who in educators' opinions is also associated with being "well-behaved".

We can therefore conclude that girls are generally perceived as "well-behaved" by teachers, whereas boys tend to be seen as "difficult". Since, in the dataset, girls were most often the eldest among siblings, eldest children were also associated with being well-behaved. This, however, may not necessarily be because the eldest child is inherently more well-behaved, but rather because the eldest child is most often a girl (in this dataset). As we can see, interpreting relationships between variables may prove challenging, even when using the market basket algorithm. One must observe the interdependence of variables carefully and draw conclusions based on the full spectrum of results rather than isolated relationships.

Third Example: Association Analysis Between the Educator's Opinion and the Parent's Opinion Regarding the Child's Behavior

The final analysis aimed to determine whether there is an association between the educator's opinion about the child's behavior and the parent's opinion. To this end, two variables were used:

a) *Educator's opinion*: level one – well-behaved child; level two – difficult child, and
b) *Upbringing difficulties*, that is, the parent's opinion of the child's behavior: level one – "very well-behaved" child; level two – "well-behaved"; level three – "it varies"; level four – "I have some problems with them"; level five – "considerable upbringing difficulties".

Table 13.4 presents a summary of the association rules, which are also visualized in Figure 13.3. The first association rule shows that if a parent assesses the child at level "2" of upbringing difficulty (a "well-behaved" child), then in 32.35% of cases, the educator also assesses the child as level "1" (a "well-behaved" child). The confidence for this rule is 64.71%, meaning that in 64.71% of cases where a parent perceives the child as "well-behaved", the educator holds a similar opinion. A correlation of 64.71% suggests a moderate association between these evaluations.

Table 13.4. Summary of Association Rules from Market Basket Analysis for the Variables Educator's Opinion – Child's Gender – Birth Order – Upbringing Difficulties

Summary of association rules
Min. support = 20.0%; Min. confidence = 50.0%; Min. correlation = 50.0%
Max. antecedent size = 10; Max. consequent size = 10

Antecedent	⇒	Consequent	Support (%)	Confidence (%)	Correlation (%)
Rather well-behaved	⇒	Educator's opinion = well-behaved	32.353	64.706	64.706
It varies	⇒	Educator's opinion = difficult	26.961	69.620	61.270
Educator's opinion = well-behaved	⇒	Rather well-behaved	32.353	64.706	64.706
Educator's opinion = difficult	⇒	It varies	26.961	53.922	61.270

The second association rule indicates that if a parent assesses the child as "it varies", then in 26.96% of cases, the educator assesses the child as "difficult". The confidence is 69.62%, meaning that in 69.62% of the cases where the parent describes the child as "it varies", the educator evaluates them as "difficult". The correlation at 61.27% suggests a moderate association between these assessments.

The third association rule shows the reverse relationship: if the educator assesses the child as "well-behaved", then in 32.35% of cases, the parent assesses the child as "rather well-behaved". The confidence for this rule is 64.71%, which means that in 64.71% of cases where the educator sees the child as "well-behaved", the parent

holds a similar view. The correlation of 64.71% confirms a moderate connection between these evaluations.

The fourth association rule indicates that if the educator evaluates the child as "difficult", then in 26.96% of cases, the parent assesses the child at the level "it varies". The confidence is 53.92%, which means that in 53.92% of the cases where the educator evaluates the child as "difficult", the parent describes them as "it varies". The correlation at 61.27% suggests a moderate association between these assessments.

In other words, one may say that parents' responses overlap to some extent with the teachers' opinions. Even if the alignment is not perfect, parents' evaluations still largely point to a certain level of upbringing difficulty when teachers describe the child's behavior as "difficult". While there are cases in the dataset that contradict this rule, the overall association rule indicates a general trend that supports trusting that most parents do indeed speak to the experience of upbringing difficulties when educators report behavioral challenges.

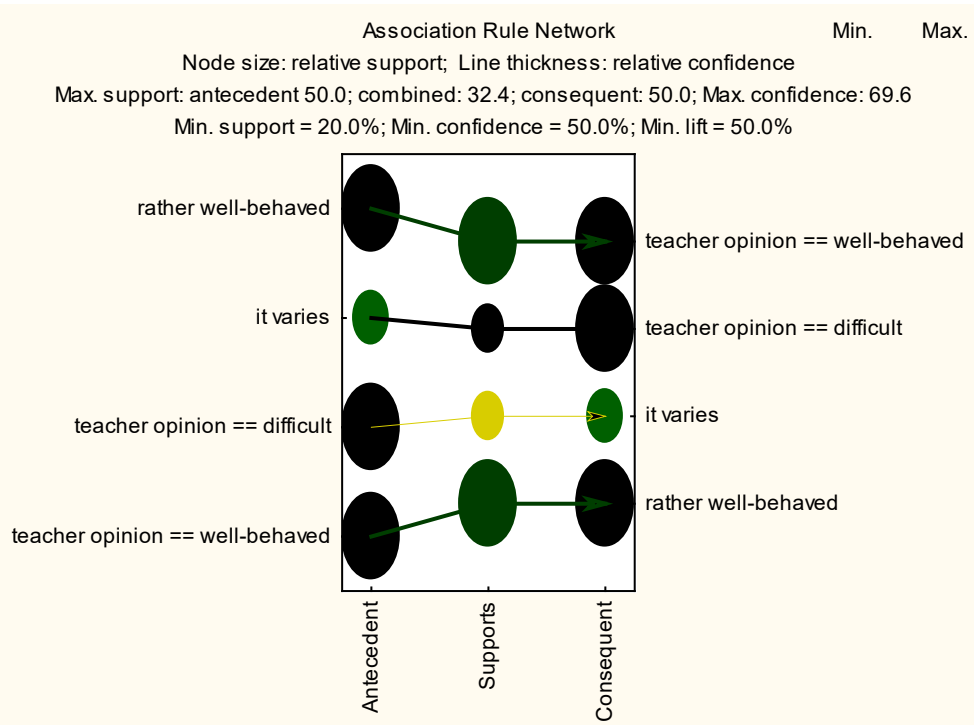


Figure 13.3. Association Rule Graph from Market Basket Analysis for the Variables Educator's Opinion – Upbringing Difficulties

The graph in Figure 13.3 presents a network of association rules between the variables "upbringing difficulty" (as assessed by parents) and "educator's opinion". The graph visualizes how often these variables co-occur (support) and how strong these associations are (confidence).

The nodes in the graph represent various levels of the variables, such as "rather well-behaved" and "it varies" for the variable *upbringing difficulty*, and "educator's opinion = well-behaved" and "educator's opinion = difficult" for the variable *educator's opinion*. The size of the nodes indicates relative support, i.e., how frequently a given level appears in the dataset. Larger nodes indicate higher support.

The edges connecting the nodes represent association rules between the variables. The thickness of the line reflects the value of confidence, that is, how often one variable level follows another. Thicker lines indicate stronger associations.

The graph analysis reveals several important patterns. First, the node representing "rather well-behaved" (upbringing difficulty) is connected to "educator's opinion = well-behaved". This rule indicates that children rated by parents as "rather well-behaved" are often rated by educators as "well-behaved". The "rather well-behaved" node is large, meaning this assessment is frequent in the dataset. The line connecting the two is thick, indicating high confidence, i.e., in many cases where the parent evaluates the child as "rather well-behaved", the educator holds a similar opinion.

Another notable pattern concerns the node "it varies" (upbringing difficulty), which is connected to "educator's opinion = difficult". This rule shows that children assessed by parents as "it varies" are often evaluated by educators as "difficult". The "it varies" node is smaller than "rather well-behaved", meaning it occurs less frequently. However, the thickness of the line still indicates a strong connection, meaning that in many cases where a parent evaluates a child as "it varies", the educator evaluates them as "difficult".

The node "educator's opinion = difficult" is connected to "it varies". This rule shows that children evaluated by educators as "difficult" are often assessed by parents as "it varies". The node "educator's opinion = difficult" is large, which means that this assessment appears frequently. The thickness of the line indicates high confidence, meaning that in many cases where the educator evaluates the child as "difficult", the parent shares a similar view.

The node "educator's opinion = well-behaved" is connected to "rather well--behaved". This rule indicates that children assessed by educators as "well-behaved" are often assessed by parents as "rather well-behaved". The thickness of the line indicates high confidence, which means that in many cases where the educator evaluates the child as "well-behaved", the parent evaluates them as "rather well-behaved".

In summary, the graph clearly shows strong associations between parental assessments of upbringing difficulty and educators' opinions about children's behavior. Children assessed by parents as "rather well-behaved" are often evaluated as "well-behaved" by educators. Conversely, children described as "it varies" by parents are frequently assessed as "difficult" by educators. These patterns may suggest a significant level of agreement between parent and educator evaluations, which can have important implications for upbringing and educational strategies.

13.2. Apriori Algorithm: Discovering Association Rules

The Apriori algorithm is one of the most important and widely used algorithms in Market Basket Analysis. Its main goal is to uncover association rules between various items in large datasets (Srinadh, 2022). These rules make it possible to identify which products or variables co-occur with high probability, which has broad applications in fields ranging from marketing to psychology. The Apriori algorithm is based on two key assumptions: the *monotonicity (anti-monotonicity) property* and the *iterative generation of frequent itemsets* (Karthiyayini & Balasubramanian, 2016). The monotonicity property states that if a given itemset is frequent, then all of its subsets must also be frequent. Conversely, if an itemset is infrequent, all of its supersets must also be infrequent. This property allows for efficient pruning of the search space. Iterative generation of frequent itemsets means that the algorithm operates in iterations, beginning with single-item sets and progressively generating larger sets until no new frequent itemsets can be found (Alawadh & Barnawi, 2022).

The Apriori algorithm consists of several steps. The first step is generating one-item candidate sets, where the algorithm identifies all individual items (products) in the dataset and calculates their support. Next, a filtering process removes the one-item sets that do not meet the user-defined minimum support threshold. The next step is generating candidate *k*-itemsets based on frequent *(k–1)*-itemsets that met the minimum support. This process is iterative. Then, the algorithm calculates the support for the *k*-item candidates by scanning all transactions in the dataset. Afterward, it filters out the *k*-itemsets that fail to meet the minimum support. Finally, the algorithm generates association rules based on the frequent itemsets that meet the minimum *confidence*.

Let us assume we have a transactional dataset from a supermarket that contains information about customer purchases. The aim is to discover which products are frequently purchased together, enabling efficient assortment management and promotional planning. The algorithm begins by generating one-itemsets and calculating the support for each product, e.g., "milk", "bread", "butter". It then filters out the products that do not meet the minimum support threshold, for example, those purchased in fewer than 5% of transactions. The next step is to generate candidate two-itemsets by forming combinations of products that meet the minimum support, such as "milk and bread", "milk and butter". The algorithm calculates the support for these two-itemsets by examining how often they appear together in transactions, then filters out the combinations that do not meet the support threshold. Finally, association rules are generated, such as "if a customer buys milk, they also buy bread", with a given support and confidence.

The Apriori algorithm has wide-ranging applications in psychology. Examples include the analysis of consumer behavior, mental health research, identification of co-occurring psychological symptoms in patients, analysis of cyberbullying phenomena, as well as educational studies—for instance, analyzing which teaching

methods are frequently used together and what impact they have on students' performance (Y. J. Chen et al., 2022; Jha & Ragha, 2013; Zainol et al., 2018).

Market Basket Analysis and the Apriori algorithm are closely related but not exactly the same. Market Basket Analysis is a technique used to uncover associations and patterns among various products or variables in large datasets. Its primary goal is to identify which products are frequently purchased together by customers. Market Basket Analysis is widely applied in marketing, retail, e-commerce, and other fields where understanding consumer behavior is crucial. The key measures used in Market Basket Analysis are support, confidence, and lift. The Apriori algorithm is one of the algorithms used to perform Market Basket Analysis. It is a specific algorithm designed for the efficient generation of frequent itemsets and for discovering association rules in large datasets. It operates iteratively, utilizing the anti-monotonicity property (if a superset is frequent, then all of its subsets must also be frequent) (Karthiyayini & Balasubramanian, 2016).

A formal description of the Apriori algorithm—including support, confidence, and lift—has been provided in detail in the chapter dedicated to Market Basket Analysis. The Apriori algorithm is a powerful tool for discovering hidden patterns in large datasets, which can lead to a better understanding of behaviors and preferences in various contexts. Its efficiency and broad applicability make it one of the fundamental algorithms in data analysis.

13.3. Association, Sequence, and Link Analysis: Extended Pattern Discovery Techniques

Association, sequence, and link analysis are advanced techniques for discovering patterns in large datasets. These methods are extensions of basic Market Basket Analysis and allow for the detection of more complex relationships between variables (STATISTICA Electronic Manual, 2012; Verma & Mehta, 2014). This chapter discusses these techniques by presenting their assumptions, operating mechanisms, and application examples, as well as comparing them with Market Basket Analysis.

Association analysis, as previously discussed, involves identifying dependencies between different elements in datasets. Algorithms such as Apriori or FP-Growth are used to detect association rules that indicate which elements frequently co-occur. In particular, association analysis is useful in the context of Market Basket Analysis, where the goal is to understand which products are frequently purchased together.

Key concepts associated with association analysis include support, confidence, and lift. *Support* measures how often a given product set appears in the entire transaction dataset. *Confidence* indicates the probability that product Y will be purchased if product X has already been purchased. *Lift* measures the strength of association between products by adjusting for the general popularity of product Y.

Sequence analysis is an extension of association analysis that takes into account the order of element occurrences. The goal of sequence analysis is to identify

sequential patterns—that is, series of events that occur in a specific order. Example applications of sequence analysis include studies of user behavior in e-commerce, where researchers analyze the steps users take before making a purchase, and research on health behavior, where sequences of events preceding specific health outcomes are examined.

Algorithms such as GSP (Generalized Sequential Pattern) and SPADE (An Efficient Algorithm for Mining Frequent Sequences) are commonly used in sequence analysis. These algorithms identify frequent sequential patterns in data using an Apriori-based approach and incorporating minimum support thresholds (Verma & Mehta, 2014).

Link analysis is a technique used to study the structure of networks and the relationships between elements in networks. It is widely used in social network analysis, where the aim is to understand how individuals are connected and what patterns of interaction exist between them. Example applications of link analysis include identifying key influential individuals in social networks and studying communication patterns in organizations.

Link analysis techniques include algorithms such as PageRank, which is used to evaluate the importance of web pages based on their connections, and community detection algorithms, which identify groups of strongly connected individuals within networks (Khan & Niazi, 2017; Stoica et al., 2024). Link analysis in social networks may be based on similarity measures between nodes (Gupta et al., 2015).

Link analysis has a wide range of applications, encompassing both social networks and other types of networks, such as web networks, criminal networks, or biological networks. SNA, in turn, is specifically focused on social networks and interpersonal relationships. Both techniques use similar mathematical and graphical tools, such as graphs, community detection algorithms, and centrality measures (e.g., degree, betweenness centrality, closeness centrality). However, techniques used in SNA are adapted to the analysis of social relationships, which may involve specialized measures and algorithms that are less commonly used in other types of link analysis.

Example applications of link analysis include the PageRank algorithm, web link analysis, criminal network analysis, and logistic network analysis. In contrast, typical applications of SNA include identifying influential individuals in social networks, analyzing organizational structures, and studying the spread of information or diseases within populations.

To summarize, link analysis and social network analysis are closely related but differ in scope and application context. Link analysis is a more general technique used to study various types of networks, while SNA is specifically focused on social relationships and social networks. Both techniques use similar mathematical and graphical tools, enabling the analysis of network structure and the identification of key patterns and nodes.

Advanced pattern discovery techniques such as association, sequence, and link analysis have wide applications across various fields. In marketing and sales,

association analysis can help understand which products are frequently purchased together, which can lead to better assortment management and promotion planning. Sequence analysis is useful in studying consumer behavior, helping to identify the steps users take before making a purchase, which can lead to the optimization of sales processes.

In health research, sequence analysis can help identify patterns of health behaviors and risk factors, which may lead to better intervention strategies. Link analysis is crucial in studies of social networks and communication, helping to identify key individuals and patterns of interaction, which can lead to a better understanding of social and organizational dynamics.

Although association, sequence, and link analysis techniques are powerful tools for discovering patterns in data, they face certain challenges. One of the main challenges is scalability, as these analyses often require processing large datasets, which can be time-consuming and resource-intensive. In addition, interpreting the results of these analyses can be difficult, as the detected patterns may be complex and multidimensional.

To summarize, association, sequence, and link analysis are advanced techniques for discovering patterns in data that have wide applications in various fields. Advanced pattern discovery techniques allow for more complex and multidimensional analyses, leading to a better understanding of behaviors and preferences. Despite the challenges related to processing large datasets and interpreting results, these techniques have great potential for further development and future applications.

13.3.1. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is an advanced statistical tool used for the analysis of sequences, where understanding hidden states influencing the observed data is crucial (Blunsom, 2004). HMM enables the modeling of systems in which observations are generated by a hidden Markov process whose states are not directly visible. This model is based on three main assumptions (Blunsom, 2004). First, the hidden Markov process has a finite number of states. Each state in the process can transition to another state with a certain probability, which is described by the state transition matrix. Second, transitions between states are described by the state transition matrix (A), where each element a_{ij} represents the probability of transition from state i to state j. Third, each hidden state generates an observation according to a specific probability distribution described by the emission matrix (B), where $b_j(o_t)$ is the probability of generating observation o_t in state j.

The HMM model is formally defined by three sets of parameters. The first is the state transition matrix (A), where $A = \{a_{ij}\}$ and a_{ij} is the probability of transition from state i to state j. The second is the emission matrix (B), where $B = \{b_j(o_t)\}$ and $b_j(o_t)$ is the probability of generating observation o_t in state j. The third element is the vector of initial probabilities (π), where $\pi = \{\pi_i\}$ and π_i is the probability that the process starts in state i (Blunsom, 2004).

In other words, the transition matrix (A) describes the trajectory of the latent feature over time, showing how it transitions between states (e.g., from point A to B to C), while the emission matrix (B) defines how this hidden feature manifests in the observed data — that is, how from each hidden state emerges the probability of occurrence of a specific indicator. The vector of initial probabilities (π), on the other hand, defines from which state the observation of the process is most likely to begin.

For those familiar with advanced methods of data analysis in psychology, the HMM model can be compared to a combination of Markov chains with confirmatory factor analysis (CFA). This model allows not only for tracking the variability of the hidden feature over time (as in classical Markov chains), but also for describing the way this feature manifests through specific observations — analogous to how latent traits in CFA are revealed through sets of indicators. The difference is that in HMM, the temporal dynamics of the process are also modelled.

The Viterbi algorithm, commonly used in the context of Hidden Markov Models (HMM), allows the determination of the most probable sequence of hidden states based on observed data (Blunsom, 2004). This algorithm is based on dynamic programming. The process begins with initialization, where the initial probabilities for each state are calculated. Then, in the recursion phase, for each observation the maximum probabilities of transition to the next states are computed. The process ends with termination, where the most probable path is selected.

The Baum-Welch algorithm is used to estimate the parameters of the HMM model — that is, the transition matrix, the emission matrix, and the initial vector — based on observational data when the sequence of hidden states is unknown (Blunsom, 2004). In other words, the Baum-Welch algorithm enables the estimation of HMM model parameters — i.e., the transition matrix between states, the emission matrix, and the initial vector — in a situation where only the observational data are available, while the sequence of hidden states remains unknown. This process involves iteratively adjusting the values of these parameters in such a way as to maximize the likelihood of the observed data, and thereby obtain the most probable structure of hidden transitions and their corresponding observations.

It is an iterative Expectation-Maximization (EM) method used to update the model parameters in such a way as to maximize the likelihood of the obtained observational data. A key element of this procedure is the forward-backward algorithm, which enables the calculation of both the probability of the entire sequence of observations and the estimation of posterior distributions of hidden states. This algorithm consists of two stages: the forward procedure, which calculates the probability of observing a given sequence up to a specific point in time, and the backward procedure, which determines the probability of continuation of this sequence from that point to the end.

For example, HMM can be used for speech recognition (Gales & Young, 2007). In such a case, the hidden states represent different phonemes (the smallest sound units), and the observations are the acoustic features of sounds recorded

by a microphone. The state transition matrix defines the probabilities of transitions between phonemes, and the emission matrix describes the probabilities of generating specific acoustic features by these phonemes. For instance, the phoneme "k" may transition to the phoneme "a" with a probability of 0.6, and the phoneme "k" may generate a certain acoustic pattern with a probability of 0.8.

Although Hidden Markov Models (HMM) are not yet widely used in psychology, their potential in the analysis of sequential data related to human behavior is promising. One example is the use of HMM to model emotional changes or behavioral patterns over time. For instance, HMM may help in the analysis of emotion sequences in patients with mood disorders, identifying hidden emotional states and predicting transitions between them based on observations such as facial expressions, voice changes, or data from biometric sensors.

Tang (2024) proposed a statistical model based on the Hidden Markov Model (HMM) to describe response processes and their differentiation among respondents. This model assumes that the response process can be modelled as a sequence of hidden states representing subsequent stages of problem solving (Tang, 2024).

In research on depression, HMM has been used to identify patterns of depression risk based on assessments of subjective well-being, coping styles, and emotion regulation strategies (Jiang et al., 2022).

In education-related research, HMM could be used to model the learning process, where hidden states represent different levels of student understanding, and observations are answers to questions or data from interactive educational platforms. This enables the dynamic adaptation of teaching strategies.

If HMM is only briefly described here, it is due to the limited application of this model in psychology to date. Most applications of HMM concern fields such as speech recognition or bioinformatics, where sequential data are more common and easier to observe. However, its potential application in psychology, particularly in the analysis of behavioral and emotional data, may gain significance with the development of technologies that record such data in real time.

To summarize, the Hidden Markov Model is a powerful tool for sequence analysis, which enables the modeling of hidden processes and temporal dependencies in data. Thanks to its versatility and wide range of applications, HMM constitutes a key element in advanced techniques of pattern discovery in data.

CHAPTER 14

Applications of the Grade Correspondence Analysis Algorithm in Data Analysis

This chapter discusses the grade algorithm, also known as Grade Correspondence Analysis or grade-based correspondence analysis (GCA, from English Grade Correspondence Analysis). Calculations using this algorithm are carried out in the GradeStat software developed by Dr. Eng. Olaf Matyja (Jarochowska, 2005a; Matyja, 2004). A detailed description of the theoretical foundations of the grade algorithm and its practical applications can be found in the monograph *Grade Models and Methods for Data Analysis* by Kowalczyk, Pleszczyńska, and Ruland (2004), published by Springer (Kowalczyk et al., 2004). This algorithm is widely used in data analysis, both qualitative and quantitative, which makes it an extremely versatile tool in scientific research. The grade algorithm is used to optimally arrange rows and columns in order to best illustrate the relationship between variables described in the columns and variables described in the rows. The algorithm's results are presented on a map of overrepresentation, which shows the strength of the relationship between rows and columns, as well as the structure of dependencies between variables.

Grade Correspondence Analysis (GCA) differs from Classical Correspondence Analysis (CA) in its approach to data transformation and interpretation. While CA focuses on the geometric representation of relationships between categories on correspondence maps, GCA takes into account both the strength and regularity of monotonic relationships. This regularity makes it possible to identify hidden patterns in the data, even when they are irregular. In contrast to CA, GCA uses indicators such as Spearman's correlation and the Gini index, and the results are presented on overrepresentation maps, which facilitates the analysis of complex relationships and trends (Szczęsny et al., 1998). The authors of the algorithm

emphasize that proper data aggregation can improve their regularity, enabling a better understanding of trends in over- and underrepresentation (Szczęsny et al., 1998).

The overrepresentation map informs about the degree of similarity between actual values in the dataset and expected values (Jarochowska, 2005a). The use of overrepresentation indicators allows for visual detection of differences between the levels of variables. Overrepresentation is defined as the ratio of observed values to expected values. When the observed and expected values are identical, the representation value equals 1. If the expected values are greater than the observed values, underrepresentation occurs (values less than 1). Conversely, if the expected values are lower than the observed values, overrepresentation occurs (values greater than 1). The expected values in a contingency table can be calculated using the following formula:

$$E_{ij} = \frac{(R_i \times C_j)}{N}$$

where E_{ij} is the expected value for cell (i,j), R_i is the sum of row i, C_j is the sum of column j, and N is the total number of observations. Calculating the expected values allows comparison with the observed values, which in turn enables the identification of overrepresentation and underrepresentation.

To illustrate the method of calculating expected values, an example observation table is presented below (Table 14.1). This table contains data on the number of people with different hair and eye colors, along with row and column sums.

Table 14.1. Contingency Table: Observed Values

	Blue Eyes	**Brown Eyes**	**Row Sum**
Blond Hair	25	35	60
Brown Hair	45	75	120
Column Sum	70	110	180

In the next table, Table 14.2, expected values for each cell have been calculated using the previously provided formula for expected values. For each cell in the table, detailed calculations are presented to illustrate how the expected values result from the proportions of the marginal sums of rows and columns.

Table 14.2. Contingency Table with Calculations: Expected Values

	Blue Eyes (C_1 = 70)	**Brown Eyes (C_2 = 110)**	**Row Sum**
Blond Hair (R_1 = 60)	E_{11} = (60·70)/180 = 23.33	E_{12} = (60·110)/180 = 36.67	60
Brown Hair (R_2 = 120)	E_{21} = (120·70)/180 = 46.67	E_{22} = (120·110)/180 = 73.33	120
Column Sum	70	110	180

As can be seen, the expected values may differ for each cell in the table, as they depend on the distribution of data in the respective row and column. For example, the expected value for the cell "Blond Hair" and "Blue Eyes" is 23.33, while for the same column but in the "Brown Hair" row, it is 46.67.

This method of calculation enables the identification of overrepresentation (observed values greater than expected) and underrepresentation (observed values smaller than expected) in the analyzed data.

After calculating the expected values, it is possible to proceed to the identification of overrepresentation and underrepresentation. This analysis allows one to understand which variable categories are particularly important in the data distribution. A supporting tool in this process is the concentration curves, which visualize the proportions between variables. Based on these results, the grade algorithm (GCA) rearranges the rows and columns, aiming to maximize the agreement in the table, which leads to a clearer representation of dependencies in the data.

The concentration curve is a graphical tool used in Grade Data Analysis (GDA) to compare the distributions of random variables. If the curve overlaps with the diagonal of the coordinate system (the 45° line), this indicates perfect proportionality between the distributions. When the curve lies above the diagonal, it indicates overrepresentation—that is, the dominance of one variable's values over the other. When it is below the diagonal, underrepresentation occurs—a deficit of values compared to the expected proportions. The concentration curve thus helps to precisely identify differences between variables and serves as the basis for further analyses, such as calculating the Gini index (Lenkiewicz, 2012).

The Gini index is the next step in the analysis, allowing for a numerical evaluation of the degree of data differentiation. Calculated as twice the area between the Lorenz curve and the square's diagonal, the Gini index summarizes inequalities in the distributions of variables. A high Gini value indicates significant differences in the data, which justifies further transformations of the table structure. Therefore, Gini serves as a key criterion that helps the grade algorithm optimize the ordering of rows and columns.

After assessing differences between distributions using the Gini index, the grade algorithm (GCA) transforms the table layout, optimizing the ordering of rows and columns. In this process, it aims to maximize concordance, e.g., through the Spearman correlation coefficient.

GCA uses the Spearman correlation coefficient (ρ) as the primary optimization indicator (Szczęsny et al., 1998). The grade algorithm arranges the rows and columns in such a way as to achieve the highest possible agreement between them (Jarochowska, 2005b). The relationships change as rows and columns are rearranged. In the case of two-dimensional tables, the algorithm seeks to maximize the Spearman correlation coefficient, which is an intuitive and statistically justified approach to data ordering (Szczęsny et al., 1998).

The overrepresentation map is a visual tool that helps to understand the extent to which different variable categories are over- or underrepresented. The interpretation of the map is based on the analysis of the arrangement of rows and columns and the representation values. Overrepresentation means that the effectiveness of a given variable level is higher than expected based on theoretical values (Jarochowska, 2005a) and is marked with warm colors (orange, red). Underrepresentation,

indicating lower-than-expected effectiveness of a given variable level, is shown with cool colors (green). Issues related to result visualization, data asymmetry, and their discretization to improve regularity have been extensively discussed in the works of Alicja Ciok, who emphasizes their importance for more accurate interpretation of data in empirical research (Ciok, 2004a, 2004b).

On the map, rows and columns that differ the most from each other are placed on opposite ends, while the most closely related variable levels are located near the diagonal. This makes it easy to identify which variable levels are most strongly associated and which show the least agreement.

The analysis of the overrepresentation map is often carried out by observing its corners. In the upper left and lower right corners are the variable levels characterized by overrepresentation, where observed values are significantly higher than expected. In contrast, the upper right and lower left corners contain variable levels with underrepresentation, where observed values are lower than expected.

The grade algorithm has wide applications not only in psychology but also in other scientific fields such as sociology, medicine, economics, and environmental sciences (Jarochowska, 2005a). Due to its ability to analyze complex datasets, this algorithm can be used for analyzing consumer behavior, market research, epidemiology, or ecology. It enables the examination of relationships between consumer preferences and various demographic features, analysis of relationships between product features and their sales in different market segments, analysis of dependencies between risk factors and the occurrence of diseases, as well as studies on the relationships between different species and their habitats.

To summarize, the grade algorithm, due to its versatility and ability to analyze various types of data, constitutes a valuable tool in scientific research. Its application in psychology, sociology, medicine, and many other fields allows for an accurate understanding of relationships between variables, which is crucial for drawing valid conclusions and making appropriate research decisions.

14.1. Grade Correspondence Analysis: Theory and Interpretation of Overrepresentation Maps

As already mentioned, grade correspondence analysis can be used to analyze qualitative data. In order to conduct an analysis for two multilevel qualitative variables, it is first necessary to construct a contingency table that will present the frequencies of co-occurrence of different levels of these variables.

In the example concerning the co-occurrence of the effectiveness of different therapies in various areas of mental health, the Generative Pretrained Transformer (GPT) language model constructed a contingency table showing how different therapies (a qualitative variable) support different aspects of mental health, such as reduction of depression, improvement of interpersonal relationships, reduction of PTSD symptoms, mood improvement, improvement of social functioning,

reduction of dissociation, reduction of anxiety, improvement of self-esteem, reduction of somatic symptoms, and reduction of psychosis (also qualitative variables). Based on general knowledge available in the literature, the GPT language model simulated values for each column, creating a contingency table that shows the frequencies of various categories of effectiveness for each therapy. These values were assumed based on the model's knowledge and understanding of the effectiveness of individual therapies in different areas of mental health.

The data contained in the table are of a simulated nature and were generated by the GPT-4 language model (OpenAI) for illustrative purposes only. They do not originate from empirical research.

The values have been presented for various therapies (CBT, Gestalt Therapy, Psychodynamic Therapy, EMDR Therapy, Systemic Therapy, Schema Therapy, Psychoanalysis, Humanistic Therapy, Ericksonian Therapy) in relation to their effectiveness in different areas of mental health. Each therapy was evaluated using a point scale, where higher scores indicate greater effectiveness.

The point values have been divided into the following categories: Very effective (24–30 points), Effective (19–23 points), Moderately effective (14–17 points), Less effective (0–13 points).

Below is a sample therapy table for Grade Correspondence Analysis. A sample therapy table for Grade Correspondence Analysis is presented below (Table 14.1).

Table 14.1. Simulated Matrix of Therapies and Aspects of Mental Health

The effectiveness categories in the table are as follows: Very effective – 24–30 points, Effective – 19–23 points, Moderately effective – 14–17 points, Less effective – 0–13 points.

Types of Therapy	Depression Reduction	Improvement of Interpersonal Relationships	PTSD Symptom Reduction	Mood Improvement	Improvement of Social Functioning	Reduction of Dissociation	Anxiety Reduction	Self-Esteem Improvement	Reduction of Somatic Symptoms	Psychosis Reduction
CBT	28	21	23	27	29	22	25	26	24	20
Gestalt Therapy	18	24	14	19	21	16	20	22	18	15
Psychodynamic Therapy	21	20	19	22	23	18	21	20	19	18
EMDR Therapy	27	25	30	29	28	26	28	27	26	25
Systemic Therapy	20	22	17	20	21	16	18	19	17	16
Schema Therapy	23	24	25	24	23	22	23	22	21	19
Psychoanalysis	19	20	18	21	22	17	19	20	18	17
Humanistic Therapy	22	23	20	24	23	21	22	23	21	19
Ericksonian Therapy	23	21	24	25	22	23	22	24	23	20

Since the table contained many variables, to facilitate analysis and visualization of the results, it was decided to shorten the names of the columns and rows. The abbreviations for the columns are as follows: depression reduction was labeled as A, improvement of interpersonal relationships as B, PTSD symptom reduction as C, mood improvement as D, improvement of social functioning as E, reduction of dissociation as F, anxiety reduction as G, self-esteem improvement as H, reduction of somatic symptoms as I, and psychosis reduction as J.

The abbreviations for the rows representing types of therapy are: CBT remains as CBT, Gestalt Therapy is labeled as TG, Psychodynamic Therapy as TP, EMDR Therapy as EMDR, Systemic Therapy as TS, Schema Therapy as TSCH, Psychoanalysis as PA, Humanistic Therapy as TH, and Ericksonian Therapy as TE.

A revised therapy table for Grade Correspondence Analysis is presented below (Table 14.2).

Table 14.2. Revised Therapy Table for Grade Correspondence Analysis

Types of Therapy	A	B	C	D	E	F	G	H	I	J
CBT	28	21	23	27	29	22	25	26	24	20
TG	18	24	14	19	21	16	20	22	18	15
TP	21	20	19	22	23	18	21	20	19	18
EMDR	27	25	30	29	28	26	28	27	26	25
TS	20	22	17	20	21	16	18	19	17	16
TSCH	23	24	25	24	23	22	23	22	21	19
PA	19	20	18	21	22	17	19	20	18	17
TH	22	23	20	24	23	21	22	23	21	19
TE	23	21	24	25	22	23	22	24	23	20

This table presents the score for each therapy in relation to different mental health criteria. It was presented to the grade algorithm, which enables the identification and visualization of relationships between columns and rows. The analysis results indicate that the Rho coefficient is 0.045065 (maximum), and the Tau coefficient is 0.030058.

In the context of our table, which presents the effectiveness of different therapies in relation to various mental health criteria, the analysis results provide valuable information about the relationships between these variables.

The Rho coefficient is a measure of the strength and direction of the association between qualitative variables in correspondence analysis. A Rho value of 0.045065 suggests that there is very little differentiation in the associations between types of therapy and their effectiveness across different areas of mental health. This value indicates that differences in the effectiveness of various therapies for particular mental health domains are negligible (this does not mean the therapies are ineffective, but rather that no significant differences were identified between therapies in treating different disorders).

The Tau coefficient is a measure of similarity between two distance matrices in correspondence analysis. A Tau value of 0.030058 suggests that there is a certain level of agreement between the expected and observed results in the context of therapy effectiveness. However, the differences between theoretical expectations and actual results are truly minimal.

The results of the grade analysis suggest that there is very little association between the columns and rows. Some therapies may be slightly more effective in specific areas of mental health than others, but these differences are nearly imperceptible. The actual results are largely consistent with theoretical expectations, as also confirmed by studies conducted by Grzesiuk and her team, although some minimal differences do exist (Szymańska et al., 2017).

The overrepresentation map (Figure 14.1) presents the relationships between types of psychotherapy and their effectiveness in various areas of mental health. By analyzing the map, we can identify a few areas of overrepresentation, which indicate a higher level of effectiveness than expected based on the expected distribution.

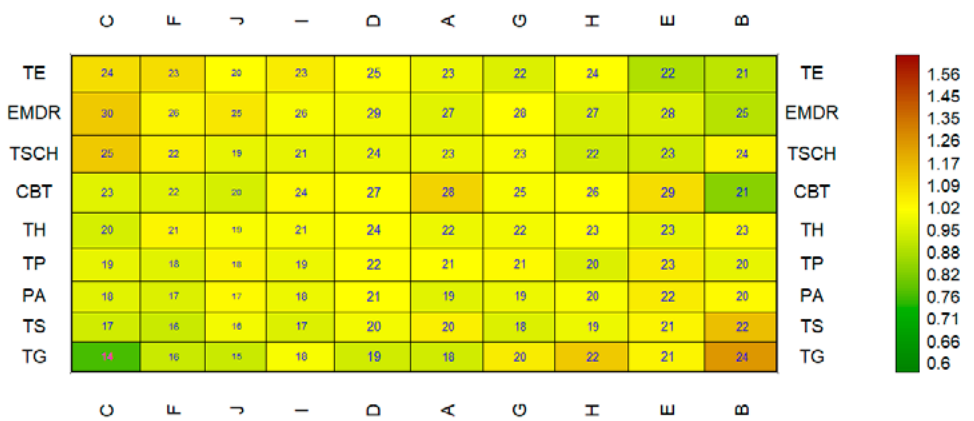


Figure 14.1. Overrepresentation Map of Therapy Effectiveness in Various Areas of Mental Health

In the upper left corner of the map, Ericksonian Therapy (TE), EMDR Therapy (EMDR), and Schema Therapy (TSCH) are located. These therapies are associated with the criteria: PTSD symptom reduction (C), reduction of dissociation (F), and psychosis reduction (J). On the map, these associations are marked in orange, indicating a certain overrepresentation. This means that Ericksonian Therapy, EMDR Therapy, and Schema Therapy are more effective in reducing PTSD symptoms, lowering dissociation, and reducing psychosis than expected.

In the lower right corner of the map, Systemic Therapy (TS) and Gestalt Therapy (TG) are located. These therapies are associated with the criteria: improvement of social functioning (E) and improvement of interpersonal relationships (B). On the map, these associations are marked in yellow, which also indicates

a certain overrepresentation, although less intense than in the upper left corner. This means that Systemic Therapy and Gestalt Therapy improve social functioning and interpersonal relationships more often than expected.

To summarize, the overrepresentation map indicates that Ericksonian Therapy, EMDR Therapy, and Schema Therapy are more effective in reducing PTSD symptoms, lowering dissociation, and reducing psychosis than expected. Meanwhile, Systemic Therapy and Gestalt Therapy improve social functioning and interpersonal relationships more often than expected. The overrepresentation map provides valuable information about the effectiveness of different therapies in specific areas of mental health.

The overrepresentation map presents the relationships between types of psychotherapy and their effectiveness in various areas of mental health. By analyzing the map, we can identify both areas of overrepresentation and underrepresentation, which indicate higher or lower levels of effectiveness than expected based on the expected distribution.

In the upper right corner of the map, Ericksonian Therapy (TE), EMDR Therapy (EMDR), and Schema Therapy (TSCH) are located. These therapies are associated with the criteria: self-esteem improvement (H), improvement of social functioning (E), and improvement of interpersonal relationships (B). On the map, these associations are marked in green, which indicates underrepresentation. This means that Ericksonian Therapy, EMDR Therapy, and Schema Therapy are less effective in improving self-esteem, social functioning, and interpersonal relationships than expected.

In the lower left corner of the map, Psychoanalysis (PA), Systemic Therapy (TS), and Gestalt Therapy (TG) are located. These therapies are associated with the criteria: PTSD symptom reduction (C), reduction of dissociation (F), and psychosis reduction (J). On the map, these associations are also marked in green, which indicates underrepresentation. This means that Psychoanalysis, Systemic Therapy, and Gestalt Therapy are less effective in reducing PTSD symptoms, lowering dissociation, and reducing psychosis than expected.

The overrepresentation map provides valuable information about the effectiveness of different therapies in specific areas of mental health, allowing for the identification of two main groups of therapies that specialize in treating different types of disorders—more than would be expected based on evenly distributed assumptions.

The first group of therapies includes Ericksonian Therapy (TE), EMDR Therapy (EMDR), and Schema Therapy (TSCH). These therapies are similar in their effectiveness at treating clinical disorders such as PTSD symptom reduction, reduction of dissociation, and psychosis reduction. The map indicates that these therapies have a specific specialization in the area of clinical disorders, which makes them more effective in treating these particular problems than other therapies.

The second group of therapies includes Psychoanalysis (PA), Systemic Therapy (TS), and Gestalt Therapy (TG). These therapies show greater effectiveness in improving social functioning and interpersonal relationships. This means that PA, TS, and TG are more effective in areas related to social and personal interactions than in treating clinical disorders.

The remaining therapies fall between these two groups, demonstrating nearly ideal representation in both areas. The overrepresentation map shows that Ericksonian Therapy, EMDR Therapy, and Schema Therapy are particularly effective in treating clinical disorders, while Psychoanalysis, Systemic Therapy, and Gestalt Therapy are more effective in improving social functioning and interpersonal relationships. This is a very interesting observation that may help mental health professionals select the most appropriate therapeutic methods for their patients based on their specific needs.

However, it should be remembered that this example comes from the predictions of the GPT language model and not from empirical research. Therefore, it should be treated more as a theoretical model than an empirical one, and with a degree of caution. Although the model provides interesting and potentially valuable insights, further empirical research is necessary to confirm these results and apply them in clinical practice.

14.2. Grade Correspondence Analysis in Structural Model Construction: The GRADSEM Approach (Grade Correspondence-Driven Structural Equation Modeling)

The GRADSEM approach (Grade Correspondence-Driven Structural Equation Modeling), presented in this chapter, is an original concept developed by Szymańska for the purpose of exploratory modeling of structural equations using grade correspondence analysis. This method is applicable in situations where no prior theoretical assumptions regarding the structure of the model exist, and where it becomes necessary to empirically determine the connections between variables. GRADSEM utilizes a grade algorithm to arrange variables in an optimal order—so that the variables most strongly related to one another are positioned along the matrix diagonal. As a result, it becomes possible to reveal the most significant relationships between rows and columns and to establish a starting point for further SEM modeling.

In this context, the grade algorithm offers unique possibilities:

a) it enables the analysis of virtually any data set, and
b) it organizes the matrix in such a way as to capture the strongest statistical structural relationships (e.g., the highest Rho values).

This arrangement of variables provides a foundation for constructing an exploratory structural model that more accurately reflects the actual distribution of dependencies contained in the data.

The algorithm allows for the computation not only of frequencies but also of other numerical values such as correlations. By inputting the data set into the GradeStat software, it is possible to perform the analysis using Spearman's rank correlation—recommended by the program's developers due to its robustness against outliers. The resulting correlation matrix may then be used to construct an overrepresentation

map, which constitutes a key stage in the process of building a structural model. The grade algorithm generates an optimal arrangement of variables in which those most strongly correlated are placed along the diagonal, which can serve as the basis for constructing the structural configuration.

In constructing a structural equation model, the researcher attempts to incorporate variables that are most strongly correlated. Although, ideally, a structural model should be grounded in theoretical assumptions, in empirical research practice this is not always feasible. In such cases, it becomes necessary to develop the model in an exploratory manner. This type of model—referred to by Jan Gajda as one that is "excavated from the data" (Gajda, 1992)—holds, however, a lower methodological status compared to a model based on the verification of theoretical assumptions, as discussed in detail in the first part of this book.

The grade algorithm offers an innovative tool for constructing structural models, particularly in cases where the researcher lacks sufficient theoretical grounding. However, the interpretation of results should take into account the limitations of exploratory models, which require further verification within a theoretical context.

When constructing a model in an exploratory fashion, the researcher is often forced to rely on intuition, since identifying the best and most optimal solution can be time-consuming and requires testing multiple combinations and different configurations in order to best align the variables. In practice, this entails checking the relationships between each variable and every other variable through simple combinatorics—a laborious and error-prone process.

In this context, the grade algorithm becomes an extremely useful tool—not only saving time but also helping to avoid potential errors in model construction. This chapter presents a method of using the grade algorithm for the construction of a structural model. First, the solution offered by the grade algorithm will be discussed, followed by the structural model built based on its results. In the next step, a theoretical solution for the same model will be presented, allowing for a comparison between the two approaches—the one derived from the grade algorithm and the one based on theory.

The software used in the analysis of structural models—such as AMOS or other SEM packages—typically provides fit indices that assist in identifying the strongest connections between variables in the saturated matrix. However, it remains the researcher's responsibility to decide which of these variables should be retained in the final model and which connections should be discarded, i.e., how to free the degrees of freedom. This process carries an inherent risk of discrepancies between the saturated matrix and the obtained matrix. If the degrees of freedom are released without accounting for the key, strongest relationships among variables, the result may be a rejected model. But how can one identify the most essential relationships in a complex network of interdependencies?

An ideal solution would be to connect every variable with every other, but such a structural model would lose its interpretive value, becoming meaningless—a model containing all possible relations explains nothing (Szymańska, 2016b).

This chapter presents the process of model construction using an example for which a theoretical solution is known. Such a choice enables a purposeful comparison of the results obtained using the grade algorithm with a solution grounded in theoretical assumptions. The grade correspondence analysis on which the exploratory structural model is based will be presented, followed by the theoretical counterpart. The chapter concludes with a comparison of both approaches, highlighting their differences and respective strengths.

As part of the analysis, a correlation study was conducted between the following variables: the discrepancy between the child's actual level of trait development and the level expected by the parent (ROZB), the degree of parenting-related difficulties experienced by the parent in the relationship with the child (TRUD), the representation of the child in the parent's mind (REPR), defense against stress (STRE), the use of pressure in the relationship with the child (PRES), constraining the child's activity (HAMO), and the parent's aggressive directiveness (AGRE). The obtained correlation values are presented in Table 14.3, which allows for the visualization and analysis of the relationships between these variables in the context of the grade algorithm. The table serves as a basis for further interpretation and the use of the results in the process of structural model construction.

Table 14.3. Correlation matrix between the variables: discrepancy (ROZB), difficulty (TRUD), representation (REPR), defense against stress (STRE), application of pressure (PRES), constraining the child's activity (HAMO), aggressive directiveness (AGRE), with correlation values

	ROZB	TRUD	REPR	STRE	PRES	HAMO	AGRE
ROZB	1.0000	0.5593	0.4150	0.4157	0.3935	0.4538	0.3195
TRUD	0.5593	1.0000	0.6014	0.7071	0.5550	0.5562	0.3505
REPR	0.4150	0.6014	1.0000	0.5387	0.4648	0.5699	0.3219
STRE	0.4157	0.7071	0.5387	1.0000	0.5524	0.5917	0.3618
PRES	0.3935	0.5550	0.4648	0.5524	1.0000	0.6261	0.3543
HAMO	0.4538	0.5562	0.5699	0.5917	0.6261	1.0000	0.5361
AGRE	0.3195	0.3505	0.3219	0.3618	0.3543	0.5361	1.0000

The analysis using the grade algorithm revealed values of Rho = 0.198189 and Tau = 0.137088, which allow for the evaluation of ordinal correspondence between the examined variables. This approach is essential for the proper construction of a structural model, as it enables the precise identification of dependencies among variables.

The grade algorithm automatically arranged the columns and rows in the table according to the optimal solution, maximizing the clarity of the presented relationships. As a result of this procedure, the most strongly related variables were placed close to each other, which facilitated the interpretation of the results. The outcome of this analysis is an overrepresentation map that graphically illustrates the

relationships obtained between variables, as shown in Figure 14.2. This map constitutes a key tool for the visualization of results and further steps in the modeling process.

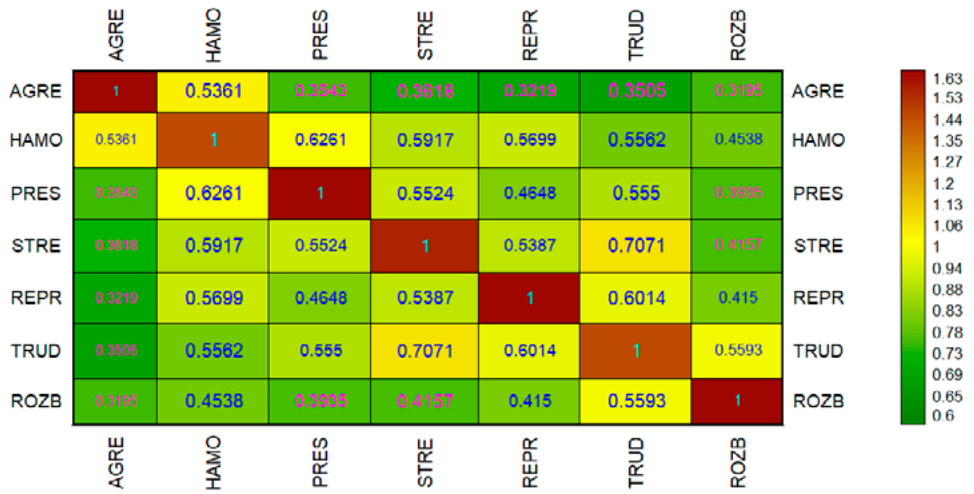


Figure 14.2. Overrepresentation map of correlations between variables related to parenting difficulties and parental responses to stress

The analysis of the overrepresentation map reveals that the variable discrepancy is most strongly associated with experienced parenting difficulty, with a correlation value of 0.559. This connection is clearly stronger than its correlations with the other variables. Accordingly, in the GRADSEM structural model presented in Figure 14.3, discrepancy was directly linked to experienced parenting difficulty. The results confirmed the strength of this relationship, indicating a value of $\beta = 0.75$.

The next variable on the diagonal of the overrepresentation map, positioned by the grade algorithm, is experienced parenting difficulty. This variable shows the strongest associations with discrepancy and with the representation of the child in the parent's mind (0.601), as well as with the stress response of withdrawal (0.707). Moreover, parenting difficulty is also related to other variables at a moderate level, such as constraining the child's activity (0.556) and the application of pressure (0.555). In the SEM model, these relationships were estimated as very strong, with values of $\beta = 0.92$ for the link with the child's representation and $\beta = 0.81$ for the link with the stress response (see Figure 14.3).

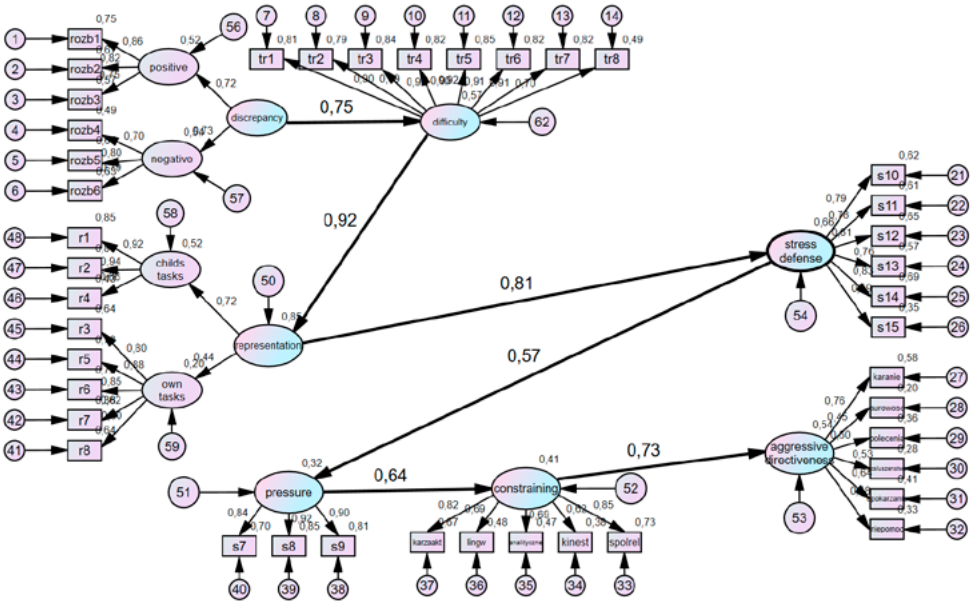


Figure 14.3. GRADSEM model generated based on the grade algorithm, showing relationships between variables related to parenting difficulties and parental responses to stress

Defense against stress shows the strongest connection with experienced parenting difficulty. However, the grade algorithm, taking into account the overall structure of relationships within the model, placed this variable on the diagonal next to the representation of the child. This arrangement reflects the dependencies identified by the algorithm and was mirrored in the structural model. The SEM results confirm the significance of this relationship, indicating a strength of $\beta = 0.81$ for the link between the representation of the child and defense against stress (Figure 14.3).

The representation of the child in the parent's mind, as the third variable in the configuration, was placed on the diagonal by the grade algorithm next to the variable defense against stress. This decision results from the optimization of the overall structure of relationships in the model, even though the child's representation is also significantly associated with constraining the child's activity (0.5699). The arrangement proposed by the algorithm was reflected in the structural model, in which the relationships between these variables were also included.

The variable defense against stress was placed by the grade algorithm in direct proximity to pressure toward the child, despite its strongest associations being with experienced parenting difficulty (0.7071) and constraining the child's activity (0.5917). This placement reflects the global optimization of relationships in the model and was replicated in the structural model, where defense against stress was linked to pressure. The SEM results confirm this connection, indicating the strength of the relationship between these variables as $\beta = 0.57$ (Figure 14.3).

Pressure was placed by the grade algorithm on the diagonal next to the variable constraining the child's activity, which reflects the strongest association for this variable (0.6261). In the structural model, this arrangement was preserved, and pressure was directly linked to constraining the child's activity. The SEM results confirm the strength of this relationship, indicating a value of $\beta = 0.64$ (Figure 14.3).

Constraining the child's activity was placed by the grade algorithm on the diagonal next to aggressive directiveness, which reflects their strongest mutual connection (0.5361). This arrangement was maintained in the structural model, where these variables were directly linked. The SEM results confirmed the strength of this relationship, indicating a value of $\beta = 0.73$ (Figure 14.3).

By applying the grade algorithm, the relationships between variables were presented as a single path running from the variable discrepancy to the variable aggressive directiveness, consisting of six relationships between variables. This arrangement was generated as optimal from the perspective of ordinal correspondence between the rows and columns of the matrix. In other words, this configuration maximizes shared variance among the variables while minimizing the impact of the weakest connections. Qualitatively, this path can be interpreted as a sequence of psychological dependencies—from discrepancy between expectations and reality, through increasing tension and pressure, to impulsive forms of reaction. Thus, the resulting structure not only enables an optimal ordering of the data but also allows for an intuitive grasp of the potential mechanism underlying the studied phenomena.

It is worth noting, however, that this approach is not entirely ideal, as it does not account for the very strong relationship between defense against stress and experienced parenting difficulty. Nevertheless, model construction always involves the need to simplify the structure. It is expected that the solution proposed by the grade algorithm compensates for these simplifications, leading to results that are as optimal as possible.

The fit of this model was analyzed in detail. The results indicate good agreement of the model with the data. Most notably, the RMSEA value is 0.065, which is below the critical threshold of 0.08 (Hair et al., 2006; Szymańska, 2016b). Additionally, the Chi^2/df ratio is 2.341, which also does not exceed the acceptable limit of 2.5. The CFI value is 0.882, which, although slightly below the desired threshold of 0.9, still allows the model to be considered a good fit due to the remaining indices such as RMSEA and Chi^2/df (Hair et al., 2006).

All relationships in the model were statistically significant, and their strength ranged from moderate to high. The strongest connection in the model reached a value of 0.92, while the weakest was 0.57. Thanks to the grade algorithm, it was possible to quickly and efficiently construct an exploratory model whose structure of relationships among variables shows a high degree of consistency with the arrangement of the empirical data, making it a reliable starting point for further structural analysis.

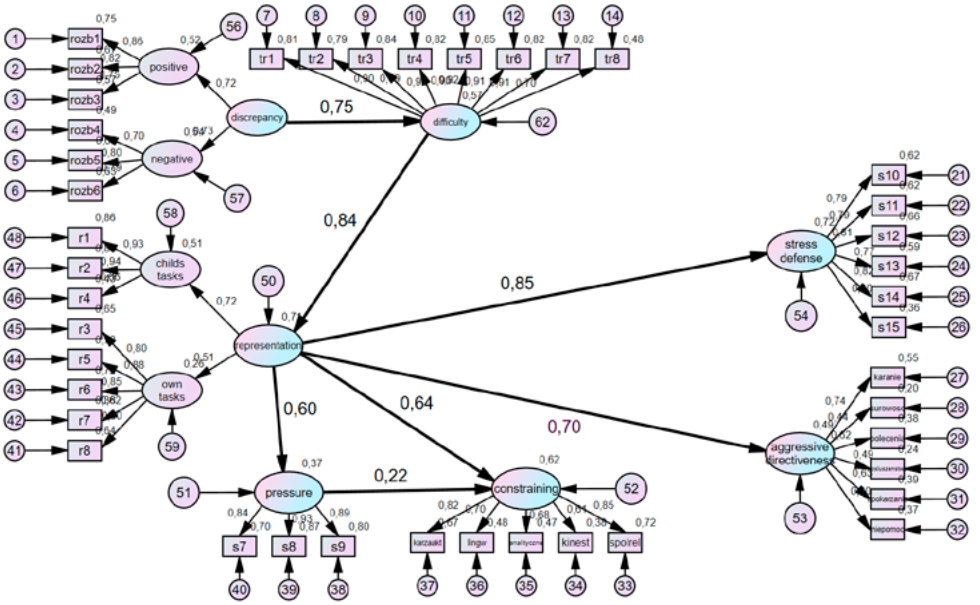


Figure 14.4. SEM model verifying theoretical relationships between variables related to parenting difficulties and parental responses to stress

When comparing the two models—the exploratory GRADSEM model created using the grade algorithm and the theoretical SEM model—both similarities and differences in their structure and results can be observed. Figure 14.4 presents the structural configuration based on theory, built on the dataset and findings presented in the book *Błąd w wychowaniu. W stronę weryfikacji teorii Antoniny Guryckiej* by Szymańska and Aranowska (Szymańska & Aranowska, 2016). A description of the sample from which the dataset used in the analyses originates is provided in Appendix C of this publication.

Both models—the exploratory and the theoretical—are similar in their main paths, such as the relationships between discrepancy, parenting difficulty, child representation, and defense against stress. Differences emerge in the further connections. The GRADSEM model (Figure 14.3) added a path from defense against stress to pressure, which was not included in the theoretical model. Both models included the link from pressure to constraining the child's activity, but GRADSEM proposed an additional relationship—from constraining the child's activity to aggressive directiveness—which was absent in the theoretical version. Conversely, the theoretical model included direct links from child representation to pressure and to aggressive directiveness, which were not incorporated in the GRADSEM model.

In comparing the strength of the paths in both models, the grade algorithm pointed to stronger connections in certain relationships, such as $\beta = 0.81$ between defense against stress and pressure, and $\beta = 0.64$ between pressure and constraining the child's activity. Meanwhile, the theoretical model focused on other connections that were somewhat weaker, yet included a greater number of direct paths.

In terms of model fit, both models achieved comparable results. The theoretical model showed better fit on several key indices: RMSEA was 0.063, which is two-thousandths lower than in the exploratory model (RMSEA = 0.065). The Chi^2/df ratio was 2.264, which is lower by 0.077 compared to the exploratory model ($Chi^2/df = 2.341$). The CFI index of the theoretical model reached 0.889, representing an increase of 0.007 compared to the exploratory model (CFI = 0.882).

It is worth noting, however, that the theoretical model included one additional path compared to the GRADSEM model, which affects the degrees of freedom and may contribute to better fit. Nevertheless, both models provided valuable information about the studied relationships, and the solution proposed by the grade algorithm demonstrated its usefulness as a method supporting the construction of exploratory models.

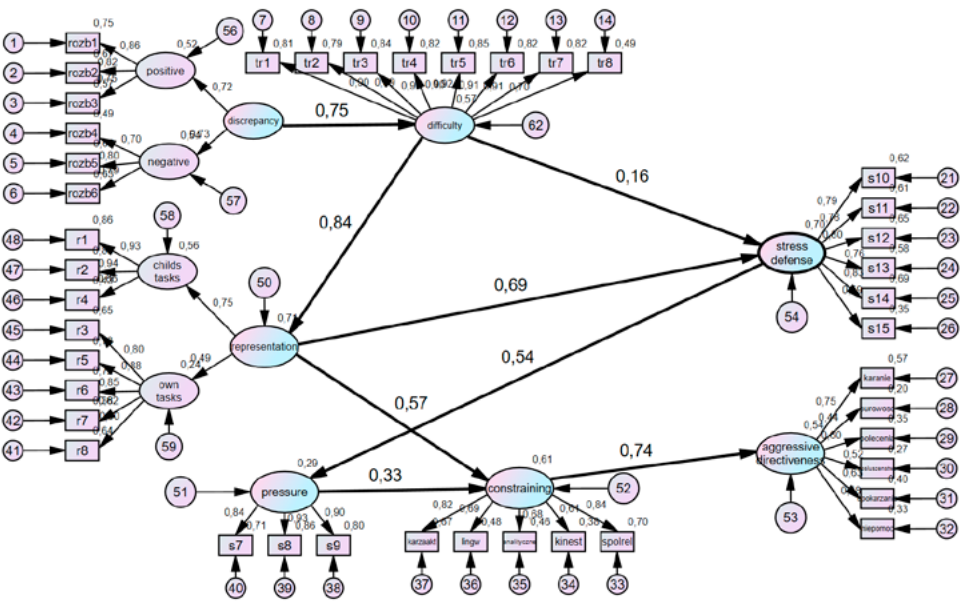


Figure 14.5. Model including other most strongly correlated variables according to the overrepresentation map, diverging from the arrangement proposed by the grade algorithm

The inclusion in the model of additional variables that, according to the overrepresentation map, appeared to be the most strongly correlated was also analyzed. Introducing such changes, however, required departing from the arrangement of rows and columns proposed by the grade algorithm, as shown in Figure 14.5. In the new model, a greater number of variables were linked, including those with the highest correlation values.

The analysis showed that this model extension resulted in only a minimal improvement in fit. The obtained index values were: RMSEA = 0.063, $Chi^2/df = 2.266$,

and CFI remained at 0.889. Despite the increased number of relationships and a lower number of degrees of freedom, the model's fit practically remained unchanged. Moreover, the increased complexity of the model significantly hindered its interpretation, making it less practical for application.

In light of these results, the model with the highest number of connections should be considered the least efficient among the three models analyzed, despite the fact that it included the theoretically strongest correlations. Ultimately, it was shown that the grade algorithm, through its ability to optimize relationships between variables, enables the construction of an exploratory structural model in a more effective and practical way.

Final conclusions

The grade algorithm proves to be an extremely useful tool in the process of constructing structural models, especially in situations where solid theoretical foundations are lacking. Thanks to its ability to analyze virtually any dataset and to optimally arrange variables, the algorithm enables researchers to identify meaningful relationships between variables that might be overlooked in traditional theory-driven approaches.

Its usefulness is multidimensional. In cases where research is conducted in new, unexplored areas, the grade algorithm can serve as an indispensable exploratory tool. It allows for the discovery of data structures and the identification of the strongest relationships between variables, which in turn leads to the formulation of new research hypotheses.

The grade algorithm can also be applied in building models that serve as alternatives to those based on existing theories. In doing so, the researcher has the opportunity to compare different models and select the one that best describes empirical reality. This approach not only enriches the research process but also provides a basis for modifying existing theories or developing new ones.

Another useful feature of the grade algorithm is its potential to support the process of verifying existing theories. Data analysis conducted with its use may reveal relationships not included in previous theoretical models. Such findings can inspire the modification, supplementation, or reconsideration of the theoretical structure in order to better reflect the empirical dependencies contained in the data.

The grade algorithm also enables the detection of hidden relationships in data that may be missed in traditional analyses. This allows for a more complete picture of the data structure, contributing to a deeper understanding of the studied phenomena.

In summary, the grade algorithm is a valuable tool in research work. Its ability to uncover optimal relationships between variables makes it an invaluable support in situations where strong theoretical foundations are lacking. At the same time, it complements traditional theory-based approaches, enabling a better understanding of the data, the discovery of new relationships, and the construction of more precise structural models.

In this way, GRADSEM—an original exploratory approach based on grade correspondence analysis—finds its application as a tool supporting the construction of structural models in conditions where clearly defined theoretical assumptions are lacking.

14.3. Verification of Circular Models: Application of Grade Correspondence Analysis

This chapter presents an original method for verifying circular models using the grade algorithm. The chapter begins with a discussion of circular models and the theories they are used to describe. It then introduces the general idea of verifying these models and the methods employed for this purpose. The next step involves a description of how the grade algorithm can be used to verify circular models. The results obtained with the help of the grade algorithm are compared with the outcomes of other methods using a real-world model as an example. Finally, the chapter presents conclusions drawn from the analyses, discusses the limitations of the reviewed methods, and demonstrates that combining them allows for more precise results.

Circular models are widely used in psychological sciences for the typologization of psychological traits such as personality or attitudes. The typologization process involves distinguishing and classifying types within a given set of defined elements. Some of the most well-known circular models include Schaefer's model of parental attitudes, Ziemska's parental attitudes wheel, Schwartz's value model, and Gurycka's circle of parental mistakes (Gurycka, 1990; Schaeffer, 1959; Schwartz et al., 2012). In the present analysis, Gurycka's circular model of parental mistakes was used as an example of applying the grade algorithm (Figure 14.6).

Typologies in circular models are based on the identification of different types and traits as well as the description of their mutual relations. Theories expressed through circular models must meet specific conditions regarding the dependencies between the distinguished elements. In this context, a key role is played by taking into account the distances between elements, which allows for the assessment of their co-occurrence and contingency.

The assumptions concerning circular models are as follows: a) elements located close to each other in the circular model are correlated, which means they co-occur, b) elements positioned on opposite sides of the circle (at an angle of 180 degrees) are negatively correlated, which means they do not co-occur, c) elements arranged at an angle of 90 degrees are uncorrelated, which suggests their random co-occurrence.

The assumptions presented above form the basis for assessing the validity of the theories expressed by circular models and for their verification. The application of the grade algorithm makes it possible to precisely verify these assumptions based on empirical data.

As noted by Lingoes: "Since we are talking here about a factor model, it is worth remembering that the points in the factor space represent the ends of vectors

originating from the coordinate system's origin, and the solution takes into account both the length of these vectors and the angles they form. In other words, $rij=h_i h_j cos\Theta_{ij}$ (the reproduced correlation between variables i and j equals the product of the lengths of the respective vectors and the cosine of the angle between them). Similarly, $a_{ik} = h_i h_k cos\Theta_{jk}$, meaning the factor loading of k on variable i equals the product of the vector lengths (where $h_k=1, k=1,2,\ldots,m$ from the model's constraints) and the cosine of the angle between the test and the factor". (Lingoes, 1977, p. 104).

Circular models not only define the placement of variables along a circle but also determine the relationships between them, using trigonometric principles to describe these dependencies. The angle between variables determines the expected correlation, with cosine values forming the basis for this description. The correlations corresponding to specific angles are as follows: a correlation of $r = 1$ corresponds to an angle of $cos = 0°$, a correlation of $r = 0$ to an angle of $cos = 90°$, and a correlation of $r = -1$ to an angle of $cos = 180°$. For example, at an angle of $cos = 45°$, the correlation is $r = 0.707$.

In a circular model with eight variables evenly distributed across a 360° angle, each variable is positioned 45° from its neighboring variables (360°/8 = 45°). In the case of a model with 10 variables, the angles between them are 36°, which corresponds to a correlation of approximately $r = 0.8$. In turn, in a model with 14 variables, the angle between them is 25°, and the corresponding correlation is r = 0.906. As the number of variables in a circular model increases, the angles between them become smaller, and the correlations between the nearest variables increase, often exceeding a value of 0.9. This makes models with a large number of variables difficult to verify empirically (Szymańska, 2025a).

The reduction of angles between variables with an increasing number of variables leads to higher correlations, which in turn decreases the accuracy and reliability of circular models. For models with more than 10 variables, it is necessary to consider whether their placement in a circular configuration is still justified. In such cases, three-dimensional models may be more appropriate, as they offer more space for variable distribution and allow for a more precise analysis of their mutual relationships.

The verification of three-dimensional models can be conducted using methods such as Support Vector Machines (SVM) with a Radial Basis Function kernel. As suggested by Szymańska (2025), such an approach offers greater possibilities for analyzing and verifying complex models of variables. The application of these methods will be discussed in the following part of this book.

Methods for Verifying Circular Models

There are several methods for verifying the assumptions underlying circular models. The two most popular are multidimensional scaling and hierarchical confirmatory factor analysis. Both methods make it possible to assess the fit of a circular model to empirical data, although they differ in analytical approach and scope of application (Szymańska, 2025a; Szymańska & Torebko, 2015).

Multidimensional scaling *"aims to construct or reconstruct a systematizing space for a specific set of elements. This space is constructed based on an ordering relationship, e.g., proximity (or distance), preference, intensity of a trait (including causal attribution, qualitative trait), affinity. In a methodological sense, multidimensional scaling itself can be interpreted as an operational definition of a given perceptual, preferential, semantic, or associative space, etc". (Biela, 1995, pp. 31–32).* Multidimensional scaling enables a spatial representation of relationships between variables, taking into account their mutual dependencies.

The fundamental mathematical assumption of this method is the transformation of data such as correlation coefficients (Δ_{ij}), representing distances or proximities between variables, into spatial data (d_{ij}), which represent these relationships in geometric form (Biela, 1995). The main goal is to obtain a spatial representation of the configuration of variables that reflects their interconnections. The mathematical aspects of multidimensional scaling have been described in detail in both international and Polish literature (Biela, 1995). This chapter presents an example of the application of this method to verify a circular model, and the obtained results are compared with those of hierarchical confirmatory factor analysis and the grade algorithm.

Hierarchical confirmatory factor analysis is another method for verifying circular models (Szymańska & Torebko, 2015). It assumes that variables located close to one another in the circular model should form hierarchical structures which, along with other variables from the same half of the circle, constitute meta-traits. Meta-traits constructed in this way should be negatively correlated with the meta-traits formed from the opposite half of the circle. A high degree of correlation consistency between variables and trigonometric assumptions, as well as good model fit, are indicators of the correctness of the circular model.

Both methods offer different approaches to verifying circular models. Multidimensional scaling focuses on the spatial representation of relationships between variables, while hierarchical confirmatory factor analysis allows for the evaluation of hierarchical structure and meta-traits. The verification of circular models requires advanced analytical tools, and combining results from various methods enables a more precise evaluation of model validity.

Verification of Circular Models Using the Grade Algorithm

As presented in Chapters 14 and 14.1, the grade algorithm allows for organizing rows and columns in such a way as to maximally highlight the relationships between them. This chapter presents the application of the grade algorithm to the correlation matrix and the cosine angle matrix in an ideal eight-element circular model. It should be emphasized that the verification of circular models using the grade algorithm is meaningful for models containing a maximum of 10 variables. With a larger number of variables, even the grade algorithm is no longer able to meaningfully separate them, and the correlation values begin to blur. In such cases, the use of three-dimensional models should be considered (Szymańska, 2025a, 2025d).

Analyzing an example of an eight-element model (Figure 14.6), the correlation between variables located closest together on the circle (at an angle of 45°) is approximately 0.707. In contrast, variables arranged at a right angle (90°) show no correlation, which is reflected in a value close to zero. Relationships between variables positioned at an obtuse angle (135°) amount to approximately –0.707, whereas correlations between variables placed opposite each other on the circle (180°) reach a value of –1.

Such visualization enables intuitive understanding of the relationships between variables in an ideal circular model, which is particularly useful in analyzing typological structures and verifying theoretical models.

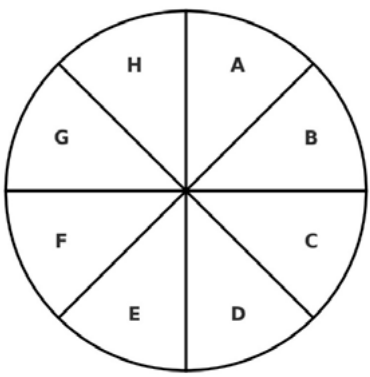


Figure 14.6. Circular model with eight theoretical dimensions.

The assumptions concerning circular models are closely related to the principles of trigonometry. Adherence to these assumptions is crucial for a typology described by a circular model to be empirically verifiable. Consequently, not all theories based on typologies can be represented in the form of a circular model, and moreover, not all models referred to as circular in the literature actually meet the required criteria. For a circular model to be considered valid, it must strictly adhere to trigonometric principles.

Table 14.4 presents the relationships between variables in a circular model composed of eight elements. According to the assumptions, the relationships on the diagonal of the correlation matrix equal one, as they reflect the association of a variable with itself. Variables located closest to each other on the circle should be positively correlated, which is marked with the symbol "+" in the table. In contrast, variables arranged at an angle of approximately 90 degrees should show no correlation, corresponding to a value close to zero. Variables located at an obtuse angle, on the other hand, should exhibit a negative correlation.

Such trigonometric assumptions form the basis for constructing valid circular models and allow for their consistent representation and verification in empirical analyses.

Table 14.4. Correlations between variables in an eight-element circular model.

	A	B	C	D	E	F	G	H
A	1	+	0	-	-	-	0	+
B	+	1	+	0	-	-	-	0
C	0	+	1	+	0	-	-	-
D	-	0	+	1	+	0	-	-
E	-	-	0	+	1	+	0	-
F	-	-	-	0	+	1	+	0
G	0	-	-	-	0	+	1	+
H	+	0	-	-	-	0	+	1

Table 14.5 presents the relationships between variables in a circular model consisting of eight variables, expressed using correlation values. According to the assumptions of circular models, variables located close to each other should be positively correlated at a high level of 0.707, which corresponds to a cosine angle of 45°. Variables positioned at an angle of 90° should show no correlation ($r = 0$), indicating their independence. In turn, variables located at an angle of 135° should be negatively correlated at a high level (–0.707). Variables placed on opposite sides of the circle, i.e., at an angle of 180°, should exhibit a maximally strong negative correlation, corresponding to a value of $r = –1$, as shown in the correlation table.

Adherence to these assumptions is essential for the validity of circular models, as it enables predictions consistent with theoretical assumptions. This allows for faithful representation of the conceptual structure and, consequently, reliable and precise empirical verification.

Table 14.5. Correlations between variables in the eight-element circular model.

	A	B	C	D	E	F	G	H
A	1	0.707	0	–0.707	–1	–0.707	0	0.707
B	0.707	1	0.707	0	–0.707	–1	–0.707	0
C	0	0.707	1	0.707	0	–0.707	–1	–0.707
D	–0.707	0	0.707	1	0.707	0	–0.707	–1
E	–1	–0.707	0	0.707	1	0.707	0	–0.707
F	–0.707	–1	–0.707	0	0.707	1	0.707	0
G	0	–0.707	–1	–0.707	0	0.707	1	0.707
H	0.707	0	–0.707	–1	–0.707	0	0.707	1

Table 14.6 presents the same relationships as Table 14.5, using cosine angles instead.

Table 14.6. Correlations from Table 14.5 Presented as Cosines of Angles Between Variables

	A	B	C	D	E	F	G	H
A	0	45	90	135	180	135	90	45
B	45	0	45	90	135	180	135	90
C	90	45	0	45	90	135	180	135
D	135	90	45	0	45	90	135	180
E	180	135	90	45	0	45	90	135
F	135	180	135	90	45	0	45	90
G	90	135	180	135	90	45	0	45
H	45	90	135	180	135	90	45	0

If we were to present the cosine angle matrix from Table 14.6 on an overrepresentation map using the method employed by the grade algorithm, it would appear as shown in Figure 14.7.

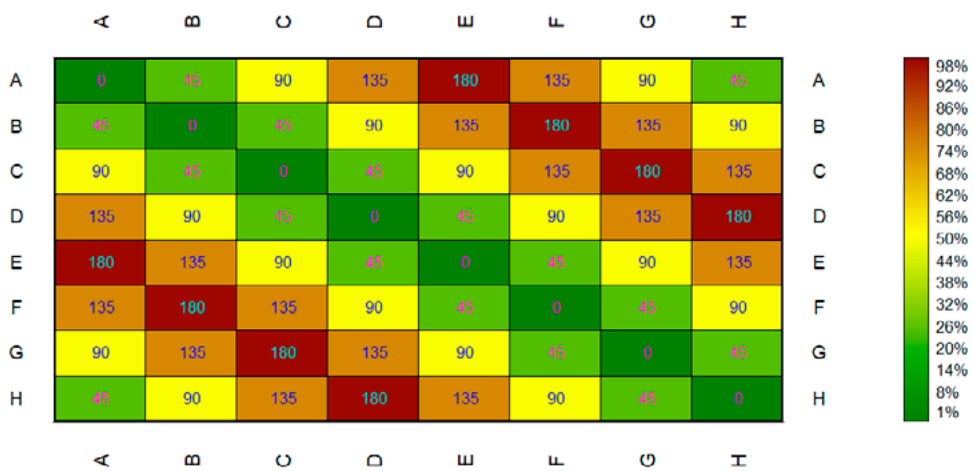


Figure 14.7. Overrepresentation map of the cosine angle matrix from Table 14.6 (generated using the grade algorithm)

Since the grade algorithm does not operate on negative values, the first step in the analysis is to transform the matrix presented in Table 14.5 in such a way that all values become positive. The simplest method is to shift the matrix values by a fixed amount—for example, by adding one to each correlation value. As a result, the value of minus one (−1) becomes zero, since (−1 + 1) = 0, and the value of minus 0.5 becomes 0.5, since (−0.5 + 1) = 0.5.

Table 14.7 presents the transformed correlation values, each increased by one, to allow the application of the grade algorithm. It is worth noting that this is not a standard correlation table, but it still accurately reflects the distances between variables,

shifted by a constant value of one. This transformation allows the original relationships to be preserved in a manner that is intuitive and relatively easy to interpret.

To facilitate the analysis, the original negative values have been marked in red and transformed into positive values. For example, the original value of 0.707 has been transformed into 1.707, whereas the value of –0.707 becomes 0.293 after transformation. This allows the reader to easily identify the original correlation values while maintaining full readability.

It is important to remember that the table pertains to relationships between variables in a circular model involving eight variables. In the case of models with a greater number of variables, these relationships may appear differently. In such situations, the angles between variables must be taken into account, and their relationships calculated according to trigonometric rules—particularly the cosine angle rule. This enables precise mapping of relationships in more complex models.

Table 14.7. Correlations Between Variables in the Circular Model After Adding a Value of 1

	A	B	C	D	E	F	G	H
A	2	1.707	1	0.293	0	0.293	1	1.707
B	1.707	2	1.707	1	0.293	0	0.293	1
C	1	1.707	2	1.707	1	0.293	0	0.293
D	0.293	1	1.707	2	1.707	1	0.293	0
E	0	0.293	1	1.707	2	1.707	1	0.293
F	0.293	0	0.293	1	1.707	2	1.707	1
G	1	0.293	0	0.293	1	1.707	2	1.707
H	1.707	1	0.293	0	0.293	1	1.707	2

The map illustrating the original order of the matrix presented in Table 14.7 is shown in Figure 14.8. It depicts the mutual relationships between the variables in the model, enabling a visual grasp of their interconnections and serving as a starting point for further analysis using the grade algorithm.

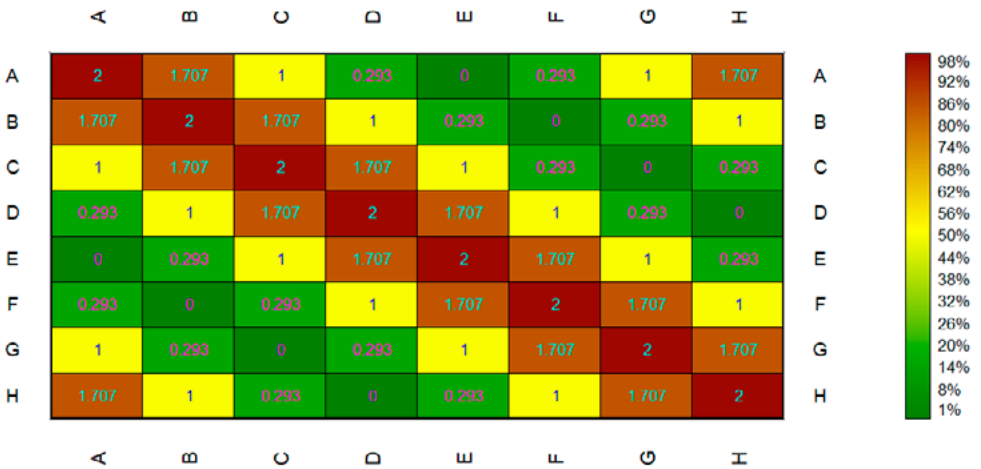


Figure 14.8. Map of the Original Order of the Matrix from Table 14.7 – Starting Point for Grade Analysis

After the matrix transformation, the data can be analysed using the grade algorithm. In the case of a perfect circular model, the result of this analysis is an overrepresentation map in which the relationships between variables are arranged in a pattern resembling a circle, as illustrated in the sample Figure 14.9.

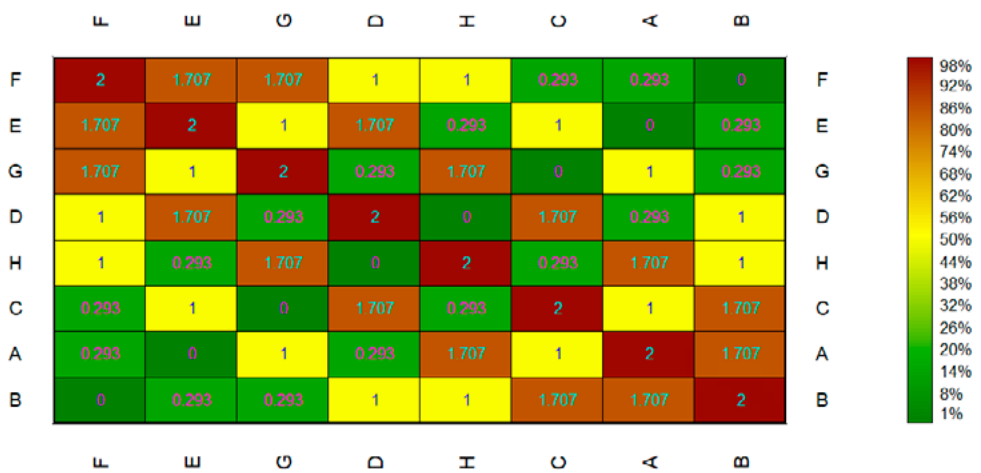


Figure 14.9. Overrepresentation Map for the Correlation Matrix Presented in Figure 14.8, After Analysis Using the Grade Algorithm (Rho = 0.486)

Following the analysis conducted with the grade algorithm, an overrepresentation map was generated, presenting the reordered columns and rows of the matrix. The diagonal of the matrix contains values of 2, representing the relationship of a variable with itself, symbolising maximum overrepresentation, and is marked in red. The ideal expected value, representing the absence of correlation, is 1. In the transformed

correlation matrix, the value of 1 corresponds to a correlation of zero, which has been noted in the map's legend.

Variables A and C were assigned a mutual angle of 90°, which means the cosine of that angle is $r = 0$. In the transformed matrix, the value of 1 reflects the ideal representation of this relationship. The overrepresentation map is characterised by a circular pattern that shows zero correlations in accordance with the expectations of the circular model. In the top-left and bottom-right corners of the map, overrepresentations marked in red and orange appear. These strongly linked variable pairs include F–E, F–G, A–B, and B–C. In contrast, the top-right and bottom-left corners of the map feature underrepresentations, which indicate the weakest connections between variables—for example, B–F, B–E, B–G, F–A, and F–C.

The diagonal of the matrix presents the strongest relationships between variables. Starting from the top-left corner of the matrix, we observe that variable F is strongly connected to E and G, while variable E shows strong connections with F and D. Similarly, variable G is closely linked to F and H, and variable D shows strong relations with E and C. In turn, variable H is strongly related to A and G, while variable C is linked to D and B. Variable A is strongly connected to B and H, and variable B to A and C.

This organisation of data in the overrepresentation map facilitates a better understanding of the patterns of relationships between variables in the circular model, providing clear information about the strength of their associations and their distribution within the model's space.

A distinctive pattern can be observed in the overrepresentation map, reflected in the arrangement of the variables. In the top-left and bottom-right corners of the map, overrepresentations highlight the most strongly connected variables. Along the diagonal running from the top-left to the bottom-right corner, double diagonals stretch across the map, further emphasising these strong relationships. Conversely, in the top-right and bottom-left corners, underrepresentations appear. Along the diagonal running from the top-right to the bottom-left corner, double diagonals indicate the weakest connections between variables.

The pattern of the overrepresentation map also includes a central circular shape representing variables that are uncorrelated with one another and located at a 90-degree angle. These variables are marked in yellow, indicating full representation corresponding to the expected correlation value of $r = 0$. Due to the applied transformation, this value has been shifted to 1 on the map. The uncorrelated variables are: a) F with D and H, b) E with G and C, c) G with E and A, d) D with F and B, e) H with F and B, f) C with E and A, g) A with G and C, h) B with D and H.

14.3.1. Theoretical Example of Verifying a Circular Model Using the Grade Algorithm

To illustrate the discussion, a theoretical example will be presented based on empirical data from a circular model consisting of eight dimensions. The data comes from

a sample of 408 individuals. The assumptions regarding the relationships between variables in this model are identical to those previously described for eight-element models, making it unnecessary to introduce any new assumptions. The relationships between variables remain consistent with the principles of any eight-element circular model. An example of such a circular model is presented in Figure 14.7.

Table 14.8 presents the correlation matrix for the dimensions described in the verified circular model. The analysis used Spearman's rank correlation, which is more resistant to outliers (Jarochowska, 2005a).

Table 14.8. Correlation matrix for the dimensions described in the circular model

	one	two	three	four	five	six	seven	eight
one	1	0.399	–0.023	–0.311	–0.503	–0.182	0.098	0.555
two	0.399	1	0.502	0.207	–0.186	–0.437	–0.334	–0.030
three	–0.023	0.502	1	0.555	0.138	–0.413	–0.626	–0.426
four	–0.311	0.207	0.555	1	0.488	–0.081	–0.387	–0.628
five	–0.503	–0.186	0.138	0.488	1	0.367	0.038	–0.474
six	–0.182	–0.437	–0.413	–0.080	0.367	1	0.559	0.103
seven	0.098	–0.334	–0.626	–0.387	0.038	0.559	1	0.466
eight	0.555	–0.030	–0.426	–0.628	–0.474	0.103	0.466	1

The same correlation matrix has been visualised as a raw data map prior to applying the grade algorithm (Figure 14.10). Around the main diagonal, the most strongly correlated variables are visible, indicating their close associations. On either side of the main diagonal, secondary diagonals extend outward, representing variables negatively correlated with the main variable on the diagonal.

When comparing the raw data map of this correlation matrix with the raw data map of the ideal circular model shown in Figure 14.8, one can observe that the current model is close to circular. A distinctive feature of the map is the orange area in the corner, indicating a strong correlation between the eighth and first variables, consistent with the assumptions of the circular model.

The colours and their distribution on the map play a key role in interpreting the results of the algorithm. Their visual representation allows for quick identification of relationships between variables and for determining the nature of the model under analysis. The colour pattern intuitively shows whether the configuration of variables corresponds to a circular model or another theoretical structure.

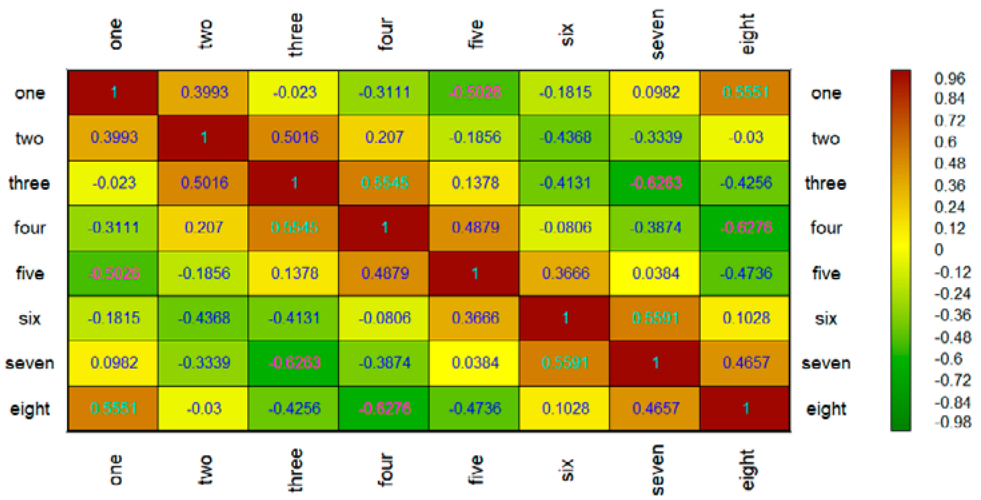


Figure 14.10. Map for the Raw Data of the Spearman Rank Correlation Matrix and for the Data Transformed by the Value of One

Since the grade algorithm cannot be computed on negative data, the correlation matrix was transformed by adding the value of one (Table 14.9).

Table 14.9. Spearman Rank Correlation Matrix Transformed by the Value of One

	one	two	three	four	five	six	seven	eight
one	2	1.399	0.977	0.689	0.497	0.818	1.098	1.555
two	1.399	2	1.502	1.207	0.814	0.563	0.666	0.970
three	0.977	1.502	2	1.555	1.138	0.587	0.374	0.574
four	0.689	1.207	1.555	2	1.488	0.919	0.613	0.372
five	0.497	0.814	1.138	1.488	2	1.367	1.038	0.526
six	0.818	0.563	0.587	0.919	1.367	2	1.559	1.103
seven	1.098	0.666	0.374	0.613	1.038	1.559	2	1.466
eight	1.555	0.970	0.574	0.372	0.526	1.103	1.466	2

Figure 14.11 presents the Spearman rank correlation matrix transformed by the value of one, whereas Figure 14.12 shows the overrepresentation map after applying the grade algorithm. The correlation value between the columns and rows is Rho = 0.335.

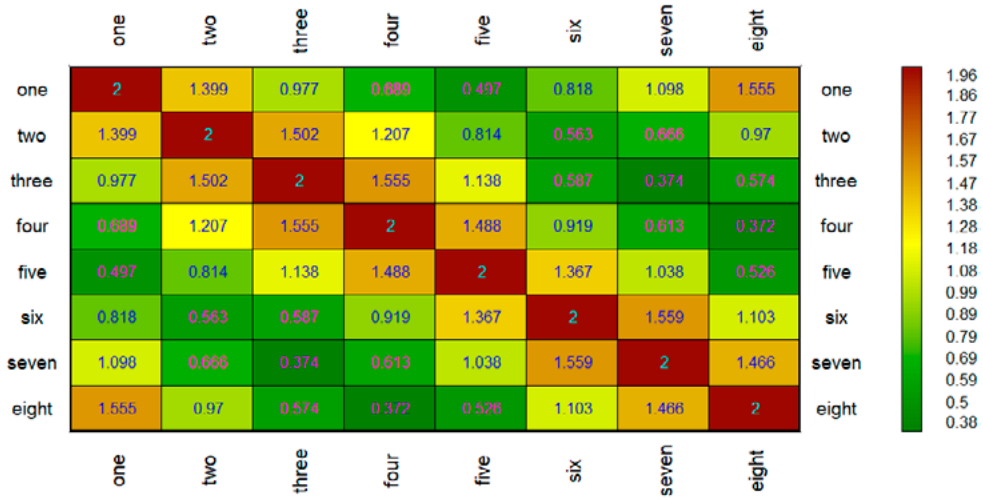


Figure 14.11. Map for the Data Transformed by the Value of One

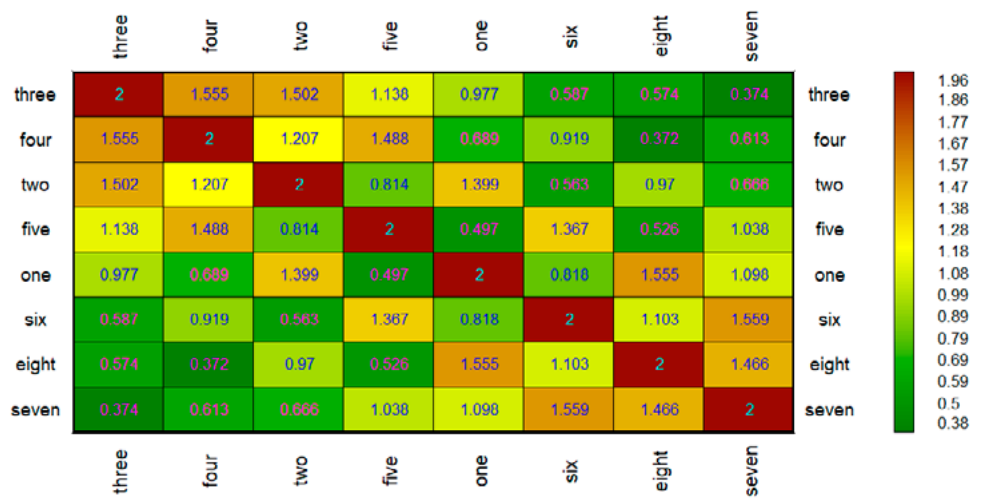


Figure 14.12. Overrepresentation Map for the Spearman Rank Correlation Matrix Transformed by the Value of One

The pattern on the overrepresentation map shown in Figure 14.12 resembles the classical pattern for the circular model, as presented in Figures 14.9 and 14.11. On both sides of the main diagonal, two sub-diagonals are visible, indicating the variables most strongly associated with those on the main diagonal. Overrepresentations occur in the top-left and bottom-right corners of the map, while underrepresentations appear in the top-right and bottom-left corners. The double diagonals extending from the top-right to the bottom-left corners reveal the weakest relationships between variables (underrepresentations). Additionally, the map

displays a circular shape indicating variables that are uncorrelated with each other and situated at a 90° angle. These variables are marked in yellow, indicating an expected value corresponding to a correlation of $r = 0$ (transformed to the value of one due to the applied transformation).

When analysing the variable *three*, it becomes apparent that it is most strongly associated with the variables *four* and *two*. In contrast, the variables *one* and *five* are not related to it, which is reflected in the 90° angle in the circular model. The variable *three* is negatively related to variables *six*, *eight*, and *seven*, with the strongest negative correlation occurring with variable *seven*, resulting from a 180° angle between them. These relationships are consistent with the theoretical assumptions of the model.

The variables *four* and *two*, which are located in the circle in the immediate vicinity of variable *three*, show positive correlations. Variables *one* and *five*, positioned at a right angle to variable *three*, are characterised by zero correlations, which also aligns with the assumptions of the model. Meanwhile, variables *six*, *seven*, and *eight*, which are arranged at an obtuse angle to variable *three*, show negative correlations, confirming the circular structure.

Variables *four* and *five* are most strongly associated with variable *three*, which results from their close proximity in the circle. The weakest correlations occur with variables *six* and *two*, which lie at a right angle to variable *four*. The negative correlations of variable *four* with *one*, *eight*, and *seven* are consistent with the circular model, as these variables form obtuse angles, with variable *eight* being positioned at 180° relative to *four*, which explains the strongest negative correlation.

Variables *one* and *three* show the strongest correlations with variable *two*, which is consistent with the circular model. The weakest correlations occur with variables *eight*, *five*, and *four*. The weak associations between variable *two* and variables *eight* and *four* result from the 90-degree angle between them. However, the correlation with variable *five* should be stronger, as the 135° angle suggests a correlation of −0.707. Variable *two* shows negative associations with *five*, *six*, and *seven*, with the strongest negative association occurring with variable *six*, which corresponds to the 180° angle between them.

Variables *five* and *four* as well as *six* show strong associations consistent with the circular model due to their close proximity. Weak correlations with variables *seven* and *three*, which are positioned at a 90-degree angle, are also consistent with theory. Meanwhile, the negative correlations of variable *five* with *two*, *one*, and *eight* mostly confirm the model assumptions, although the correlation with variable *two* is lower than expected. The strongest negative correlation with *one* is consistent with the obtuse 180-degree angle.

Variables *one*, *two*, and *eight* show strong correlations, consistent with close neighbourhood in the circle. Weak associations with variables *three* and *seven*, resulting from the 90-degree angle, also correspond to the model. Negative correlations with *four*, *five*, and *six* mostly align with expectations, although the correlation between variables *one* and *six* is weaker than expected for the 135° angle.

Variables *six*, *five*, and *seven* are strongly correlated, which results from their close proximity. The relationships between *four* and *eight* indicate a lack of correlation, consistent with the 90-degree angle. Negative associations of variable *six* with *two*, *three*, and *one* confirm the model assumptions, although the correlation with *one* is weaker than would be expected from the 135° angle.

Variables *eight*, *seven*, and *one* show strong correlations consistent with the model. The lack of associations with *two* and *six*, resulting from the 90-degree angle, confirms the assumptions. Negative relationships with *four*, *three*, and *five* are also consistent with the model, especially the strong negative association with *four*, resulting from the 180° angle.

Variables *seven*, *six*, and *eight* are strongly correlated in line with the model. Weak correlations with *five* and *one*, positioned at a 90-degree angle, are consistent with theory. Negative associations with *three*, *two*, and *four* also confirm the model assumptions, with the strongest negative correlation with *three* resulting from the 180° angle.

In summary, the circular model shows consistency with its assumptions in the relationships between closely positioned variables and in the case of variables arranged at a 90-degree angle. Deviations concern the correlations between variables positioned at a 135-degree angle and, in some cases, between opposite variables. These discrepancies may result from empirical data limitations or the influence of outliers. Further research should take into account the possibility of improving consistency by exercising greater control over data quality or by adjusting the verification algorithm.

This model was subsequently verified using multidimensional scaling, one of the popular methods, to check whether the conclusions obtained would be similar. The multidimensional scaling solution, presented in Figure 14.13, shows that the arrangement of variables resembles the circular model. The coordinates of the variables are provided in Table 14.10. Variables that are located close to each other in the theoretical model are also positioned closely in the multidimensional scaling solution.

Moreover, variables that are positioned opposite each other in the circular model (theoretically assumed to be opposites) are also placed opposite one another in the multidimensional scaling solution and form a 180-degree angle. This applies to the following variable pairs: *one – five*, *two – six*, *three – seven*, *four – eight*. Variables that are arranged at a 135° angle in the circular model also form an obtuse angle in the multidimensional scaling solution. These include the pairs: *one – four*, *one – six*, *two – seven*, *two – five*, *three – six*, *three – eight*, *four – one*, *four – seven*.

Variables arranged at a 90° angle in the circular model are found at distances corresponding to that angle in the multidimensional scaling solution. This concerns the following variable pairs: *one – three*, *one – seven*, *two – four*, *two – eight*, *three – five*, *four – six*, *five – seven*, and *six – eight*. Variables that are adjacent in the circular model are also located close to one another in the multidimensional scaling solution, confirming the agreement between the multidimensional scaling solution and the theoretical model.

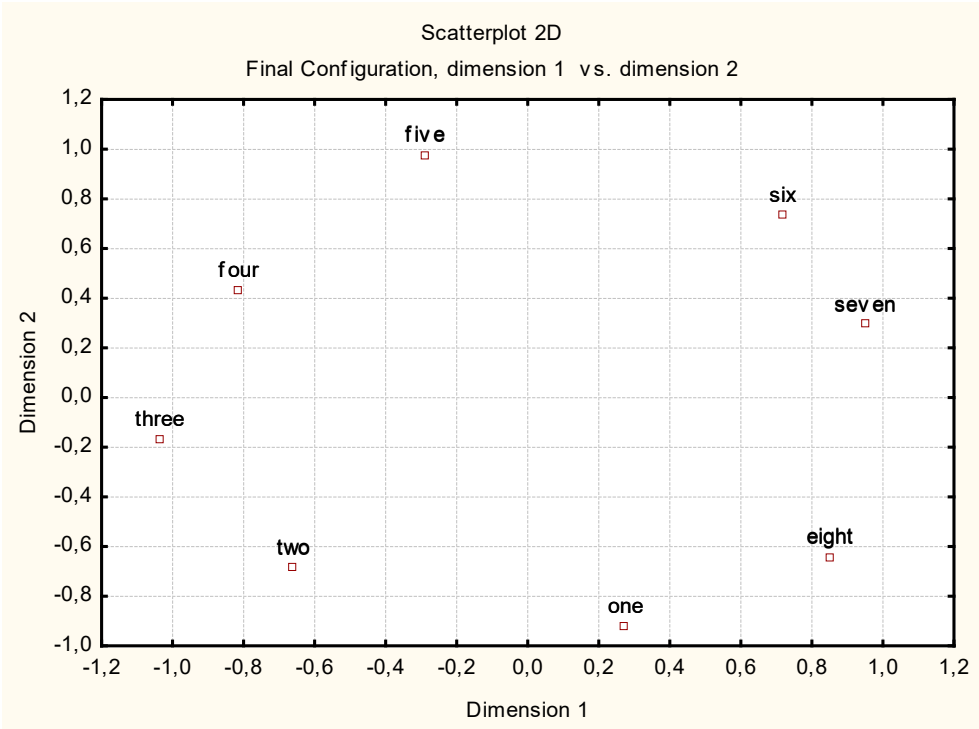


Figure 14.13. Results of Multidimensional Scaling for Variables in the Circular Model from One to Eight

The multidimensional scaling solution suggests that the model described is very close to circular. However, based on the results of the grade algorithm, this solution should be approached with greater caution. The grade algorithm confirmed the relationships between the variables positioned closest to each other in the circular model, which was reflected in the multidimensional scaling solution. These relationships were accurately represented, which strengthens the credibility of both the algorithm and the multidimensional scaling method in the context of nearest-neighbour variable positioning.

The relationships between variables arranged at a 90° angle in the circular model were also confirmed by the grade algorithm. As expected, these associations were close to zero, which applies to the variable pairs: *one – three*, *one – seven*, *two – four*, *two – eight*, *three – five*, *four – six*, *five – seven*, and *six – eight*. These consistencies indicate the accuracy of the representation in the multidimensional scaling solution.

Unfortunately, the grade algorithm also revealed significant inconsistencies with the assumptions of the circular model for other variable relationships involving variables positioned farther apart in the circle. This especially concerns variables that should be negatively correlated or show distinct differences in correlations due to their placement in the model. Such discrepancies suggest that although the multidimensional scaling solution resembles the circular model, not all theoretical assumptions are met, which requires further analysis and verification.

Table 14.10. Coordinates of Variables According to Multidimensional Scaling

Variable	DIM. 1	DIM. 2
one	0.27648	−0.922945
two	−0.65999	−0.688206
three	−1.03562	−0.170809
four	−0.81403	0.428932
five	−0.2889	0.974594
six	0.72029	0.731366
seven	0.95005	0.290905
eight	0.85172	−0.643835

First, the variables positioned opposite each other at a 180-degree angle were negatively correlated, but only at a moderate level. This concerned the relationships between the following variable pairs: *one – five*, *two – six*, *three – seven*, *four – eight*. Expecting these relationships to be maximal, i.e., equal to −1, is unrealistic due to measurement error. Nevertheless, assuming these variables are conceptual opposites, one would expect stronger correlations than −0.707. The moderate associations between these variables seriously undermine the circularity of the model under analysis.

Second, the grade algorithm showed that the relationships between variables arranged at a 135° angle are too weak. This concerns the associations between the following variables: *one – four*, *one – six*, *two – seven*, *two – five*, *three – six*, *three – eight*, *four – one*, *four – seven*. These weak associations also fail to meet the assumptions of the circular model.

Multidimensional scaling (MDS) suggested that the distances between variables were well preserved, but this was not the case in reality. Its main limitation lies in the necessity of projecting the results onto a two-dimensional plane, which leads to a flattening of the data structure. While preserving the angles between selected variables (e.g., ≈45°), the algorithm arranges the points in a circle-like layout—this is a projection artefact rather than a feature of the actual configuration. Meanwhile, the grade algorithm is not subject to such limitations, making its solution more precise and accurate. The algorithm enables analysis of the relationship between each variable and every other. The overrepresentation map further allows an assessment of how closely the solution resembles a circular arrangement.

Hierarchical confirmatory factor analysis was applied as the next method for verifying circular models. The relationships between variables in circular models reflect not only cosine angle measures but also the theoretical axes of the model. It is assumed that variables assigned to these axes form meta-traits. For example, variables *one*, *two*, *three*, and *four* were assigned to one meta-trait, while variables *five*, *six*, *seven*, and *eight* were assigned to the other. These meta-traits should be negatively correlated. Practically every circular model assumes the existence of at least one theoretical axis, and some models contain several. The next part of the chapter will show how these axes can be used to verify the circular model.

The results of hierarchical factor analysis were compared with the outcomes obtained through multidimensional scaling and the grade algorithm. The assumption was made that the model contains one dominant axis dividing the variables into two groups. These groups were labelled as "plus" and "minus" traits. In the subsequent interpretation, the "plus" traits (including variables *one*, *two*, *three*, and *four*) were considered to represent assertive attitudes, whereas the "minus" traits (*five*, *six*, *seven*, and *eight*) reflected conciliatory attitudes. The analysis yielded a hierarchical structure presented in Figure 14.14.

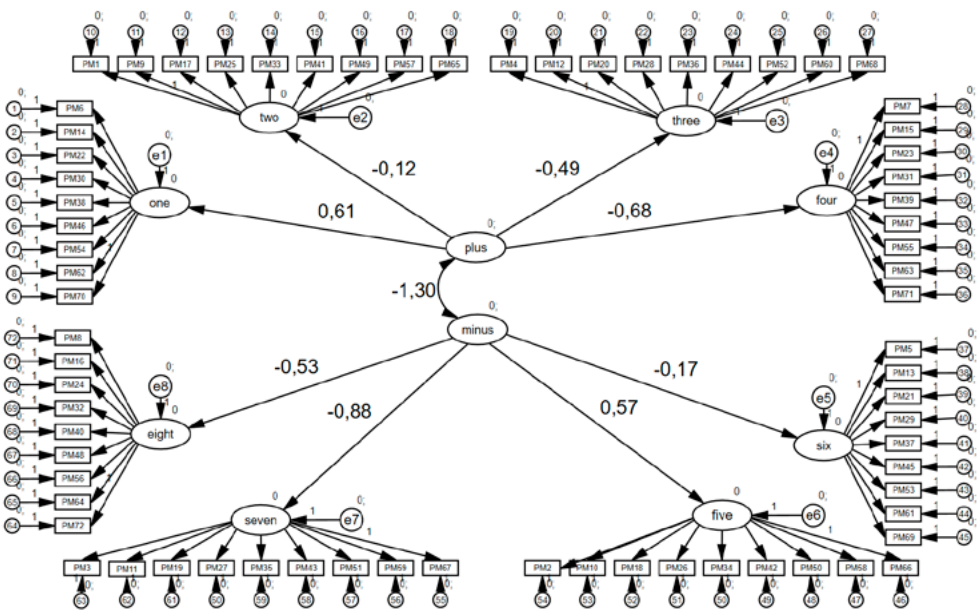


Figure 14.14. Solution of Hierarchical Confirmatory Factor Analysis for the Circular Model Based on a Single Axis

As can be seen in Figure 14.14, the assumptions of the theoretical circular model were violated. Variables *two*, *three*, and *four*, which according to the adopted convention should belong to the "plus" group (interpreted as assertive attitudes), show negative correlations with the structure of that group. A similar inconsistency concerns variables *six*, *seven*, and *eight*, assigned to the "minus" group (interpreted as conciliatory attitudes), which also correlate negatively with their own pole. As a result, both overarching structures—"plus" and "minus"—remain negatively related to one another, which may suggest that instead of coherent poles, we are dealing with an internally inconsistent or reversed pattern of behaviour.

Exceeding the value of one in the case of some associations may result from the fact that the factor loadings of variables are both positively and negatively correlated within the overarching structures. It is also worth noting that this is not a standard correlation. As emphasised by Jöreskog, all relationships between variables in

structural models are in fact regressions, which can exceed the value of one (Jöreskog, 1999). The principle of restricting values to the range from −1 to 1 applies only to cases where orthogonalisation has been used—in this case, the solution was not orthogonalised, and therefore exceeding this limit is permissible.

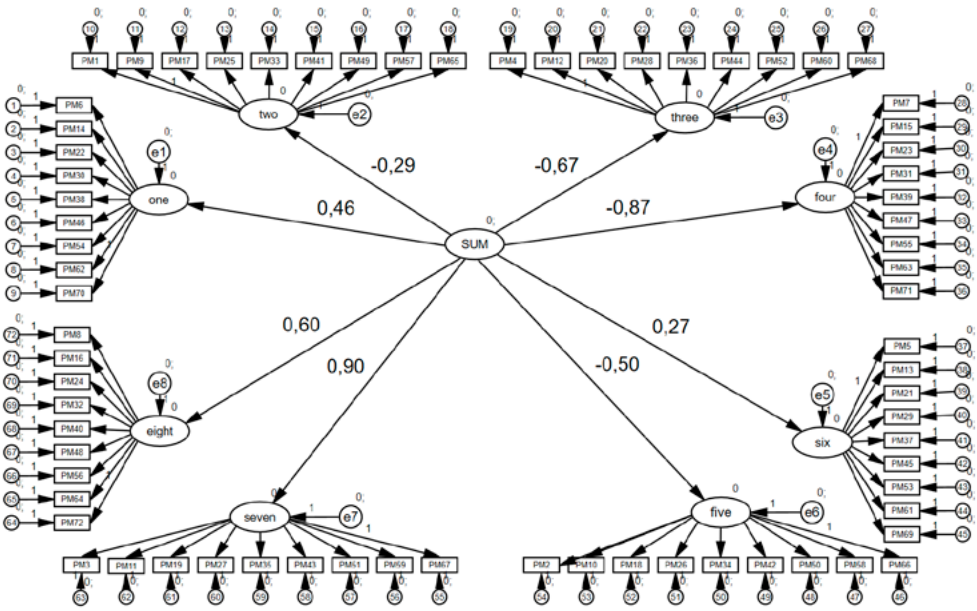


Figure 14.15. Solution of Hierarchical Confirmatory Factor Analysis for the Circular Model Based on the Existence of a Single Factor

The variables were included in a single structure, which revealed that variable *one*, which theoretically should correlate positively with variables *two*, *three*, and *four*, is in fact negatively associated with them. Similarly, variable *five*, which should show positive correlations with variables *six*, *seven*, and *eight*, also demonstrates negative associations. Figure 14.15 presents the relationships between variables that, in the circular model, should lie opposite each other. These relationships were expected to approach the value of one, but in the case of variables *two* and *six*, and *one* and *five*, their associations are too low to be considered opposites. This clearly indicates that variables *one* and *five* do not belong to their respective theoretical variable groups, which undermines the coherence of the model.

As shown in the overrepresentation map (Figure 14.12), variable *five* is most strongly associated with variable *four*, which belongs to the opposite structure. Similarly, variable *one* shows its strongest association with variable *eight*, also part of the opposite structure. The results obtained using the grade algorithm and hierarchical confirmatory factor analysis indicate serious concerns regarding the circular nature of the model under investigation. Both methods revealed inconsistencies that undermine the theoretical assumptions of the model.

In contrast, multidimensional scaling did not provide any new or significant insights into this discussion. Although it suggested that the distances between variables were appropriately preserved, it did not confirm the findings obtained via the grade algorithm and confirmatory factor analysis. Multidimensional scaling turned out to be insufficient and inadequate for evaluating the circular nature of the model. The results of this method may be misleading or incomplete in the context of more advanced analytical techniques.

Given the contradictory results obtained using different methods of model verification, a methodological discrepancy becomes apparent. The grade algorithm and confirmatory factor analysis indicate problems with the circular nature of the model, while multidimensional scaling suggests that the model aligns with the assumptions. This discrepancy highlights the limitations of individual methods and the need to use them complementarily to obtain a more comprehensive picture of the model.

Instead of relying solely on one method, it is advisable to apply them in a complementary manner. The results of multidimensional scaling may serve as a general assessment of model fit, whereas the grade algorithm and confirmatory factor analysis can provide more detailed information about potential issues. The results of multidimensional scaling, in the context of this analysis, should be interpreted with caution, and not used as the sole tool for evaluating the circular model. Further research and analysis are needed to better understand the limitations and strengths of each method and to develop a more integrated approach to verifying circular models.

14.3.2. Example of Verification of the Circular Model of Parental Mistakes Using Grade Correspondence Analysis

Let us now examine how the circular model was verified using the analysis of parental mistakes committed by parents. The study involved 402 women who answered questions about the parental mistakes of their own parents during childhood. The nature of the study was retrospective, which allowed participants to reflect on the past and assess their upbringing experiences from a temporal distance.

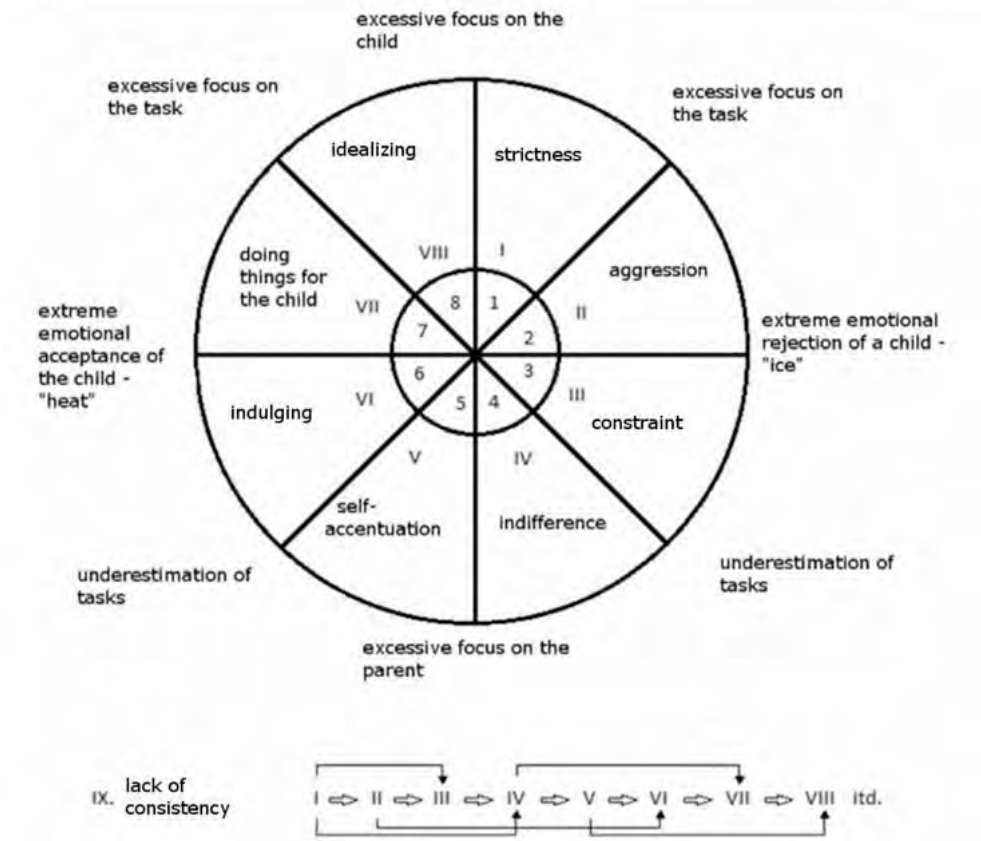


Figure 14.16. Circular Model of Parental Mistakes by Antonina Gurycka

The circular model of parental mistakes by Antonina Gurycka, presented in Figure 14.16, describes eight primary parental mistakes divided into three key dimensions: warm and cold mistakes, mistakes resulting from a focus on the parent's own tasks, and mistakes related to focusing on the child and their tasks. Each mistake can be assigned to two axes: the axis of warm or cold mistakes, and the axis indicating whether parental attention is focused on the parent and their duties or on the child and their needs.

Cold mistakes focused on the child include *rigorism* and *aggression*. Rigorism manifests as excessive demands and strictness toward the child, while aggression takes the form of violence—both physical and verbal. Other cold mistakes, this time focused on the parent, include *constraining activity* and *indifference*. Constraining activity involves limiting the child's actions and suppressing their expression, whereas indifference indicates a lack of emotional involvement in the relationship with the child.

Warm parental mistakes, resulting from a focus on the parent, include self-accentuation and indulgence (helplessness). *Self-accentuation* involves an excessive

focus on one's own needs at the expense of the child, whereas *indulgence* refers to a lack of assertiveness and consistency in upbringing. Within the group of warm mistakes centred on the child, we find doing things for the child and idealisation. *Doing things for the child* entails excessive assistance and completing tasks on the child's behalf, while *idealisation* consists in overestimating the child and perceiving their abilities and behaviours in an unrealistic manner.

All these dimensions of parental mistakes are interconnected, allowing for the observation of their mutual relationships. Cold mistakes reflect a distanced, emotionally detached approach, while warm mistakes point to excessive emotional involvement. The distinction between focus on the parent's tasks and the child's tasks highlights where parental attention is directed—towards one's own needs or those of the child. The circle of parental mistakes illustrates how these various errors may reinforce one another, creating a cycle of unfavourable behaviours in the parenting process.

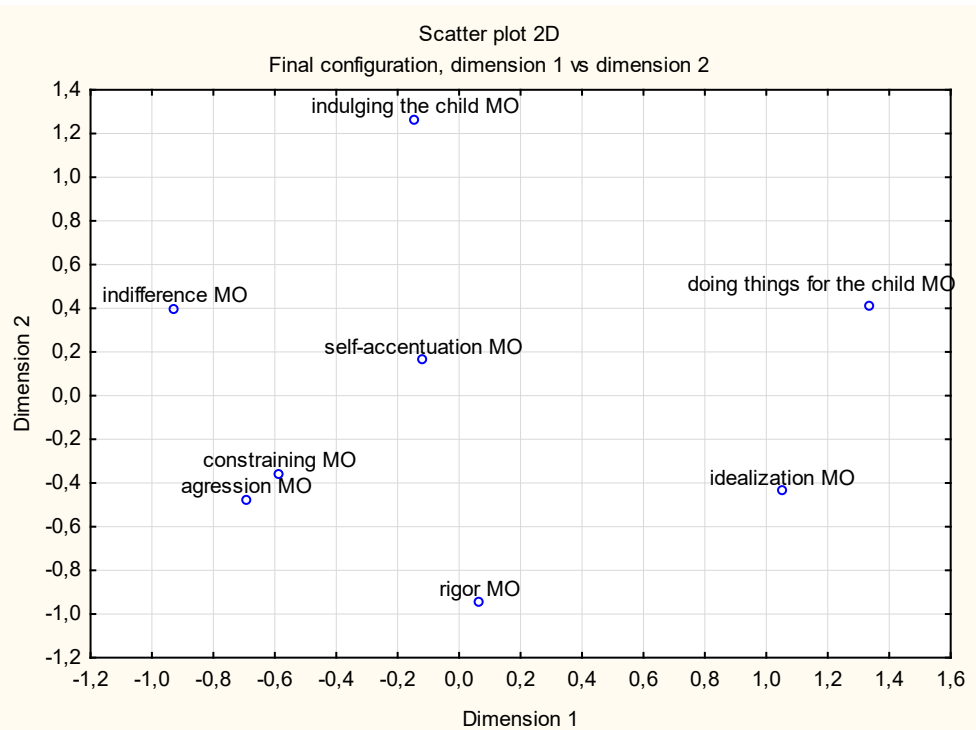


Figure 14.17. Multidimensional Scaling Solution

Based on the results of the multidimensional scaling (Figure 14.17), it can be observed that parental mistakes have been organised into a circular configuration. The mistakes are positioned close to one another, which aligns with theoretical assumptions. For example, *rigour* is located between *idealisation* and *aggression*, at equal distances from both, which corresponds to theoretical predictions. Near

idealisation is the mistake of *doing things for the child*, which also fits the assumptions of the model. Four mistakes—*aggression*, *rigour*, *idealisation*, and *doing things for the child*—fully reflect the theoretical assumptions.

The mistake of *indulgence* is situated between *doing things for the child* and *indifference*, which also matches the theory. However, the mistake of *self-accentuation*, which was theoretically expected to be positioned closest to *indulgence* and *indifference*, was located centrally on the plane, suggesting its similarity to all mistakes. This placement diverges from theoretical assumptions. The remaining mistakes, such as *indifference*, were arranged in accordance with predictions. However, *aggression* and *constraining* activity appeared too close to one another, which may suggest an excessive correlation between them. Fortunately, this correlation does not reach a value of one, which means they may still be treated as distinct categories.

In summary, the results of the multidimensional scaling presented in Figure 14.17 indicate that aside from the location of *self-accentuation* and the close proximity of *aggression* and *constraining* activity, the positioning of the remaining mistakes is consistent with the assumptions of the circular model. The circular arrangement of mistakes and their distances correspond to Gurycka's theory. Moreover, the adjacency of mistakes remains coherent with theoretical predictions. This model may be considered circle-like, although further reflection is needed regarding the position of *self-accentuation* in relation to other parental mistakes, as the theoretical foundations of this relationship have not been sufficiently described. The coordinates of the variables in the circular model of parental mistakes based on multidimensional scaling are presented in Table 14.11.

Table 14.11. Coordinates of variables in the circular model of parental mistakes according to multidimensional scaling

Mistake	Dimension 1	Dimension 2
Rigour	0.067792	–0.948844
Aggression	–0.692178	–0.479524
Constraining	–0.583435	–0.358835
Indifference	–0.924597	0.395796
Self-accentuation	–0.115441	0.160123
Indulgence	–0.141969	1.260417
Doing things for the child	1.337642	0.410155
Idealisation	1.052185	–0.439288

To validate the circular models, a second method was applied—hierarchical confirmatory factor analysis (Szymańska & Torebko, 2015). This method is based on the assumption that variables located along the same dimensions of the circle should form hierarchical, that is, multi-level, structures. Within the theory of parental mistakes proposed by Gurycka, *rigour* and *aggression* were assigned to one hierarchical structure; *constraining* activity and *indifference* to the second; *self-accentuation* and

indulgence to the third; and *doing things for the child* and *idealisation* to the fourth. This division stems from the theory in which parental mistakes are described along two dimensions: warm vs cold, and focus on the parent's tasks vs focus on the child's tasks.

These hierarchical structures were subsequently grouped into two overarching categories: warm mistakes and cold mistakes. Since in Gurycka's model warm and cold mistakes are located on opposite sides of the circle, it was assumed that they are negatively correlated.

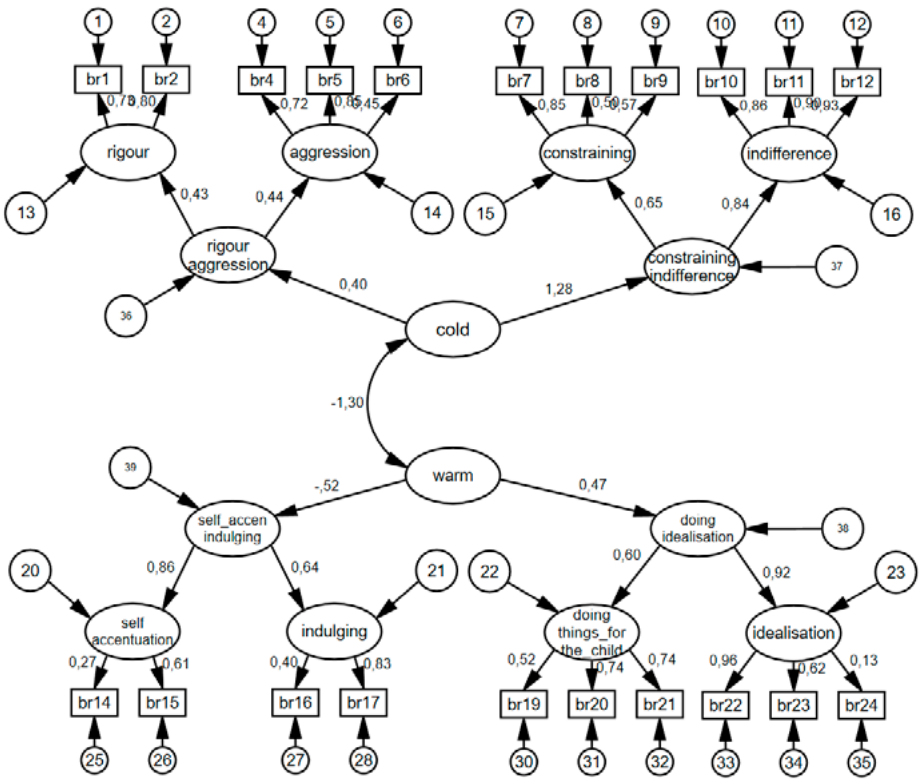


Figure 14.18. Confirmatory factor analysis of the circular model of parental mistakes according to Gurycka's theory

Figure 14.18 presents the results of confirmatory factor analysis aimed at verifying the accuracy of assigning variables to the structures described in the circular model. The analysis was based on data from the sample described in Appendix A. The primary goal of this modeling was not merely the visualisation of distances between variables, but rather the evaluation of the model's consistency with theoretical assumptions concerning the organisation of variables within the circle. If the circular model were correct, the variables belonging to the same structures should correlate positively with their assigned factors, while warm and cold mistakes should correlate negatively with one another.

However, the presented solution not only fails to fit the data (CFI = 0.762, RMSEA = 0.106), but also demonstrates that although *rigour* and *aggression*, as well as *constraining* and *indifference*, correlate meaningfully with the cold mistakes structure, *self-accentuation* and *indulgence* also correlate positively with this structure, which contradicts theoretical assumptions (Figure 14.20). Furthermore, *doing things for the child* and *idealisation*, which theoretically should correlate positively with *self-accentuation* and *indulgence* and form a cohesive warm mistakes structure, show a negative correlation with that structure. In other words, *doing things for the child* and *idealisation* break away from their group, becoming entirely unconnected to the remaining parental mistakes, even correlating negatively with them. Meanwhile, all other mistakes remain positively interrelated. This indicates a violation of the assumptions of Gurycka's circular model.

Interestingly, this problem is not as clearly visible in the results of multidimensional scaling. Except for *self-accentuation*, which was positioned at the centre of the circle, the remaining mistakes appear to be arranged according to the assumptions. However, it should be noted that *indulgence* and *self-accentuation* were located worryingly close to the cold mistakes and show correlations that contradict the assumptions of Gurycka's circular model.

To better illustrate these relationships, another factor analysis was conducted. This time, all parental mistakes were reduced to one overarching variable—*parental mistakes* (Figure 14.19). The results of this analysis show that *constraining* and *indifference*, *rigour* and *aggression*, as well as *self-accentuation* and *indulgence* are positively correlated with the entire structure, whereas *doing things for the child* and *idealisation* show a negative correlation. Although these relationships are less clearly reflected in the results of multidimensional scaling, the results of factor analysis unequivocally indicate that the current form of the model does not meet the theoretical assumptions and cannot be regarded as correct.

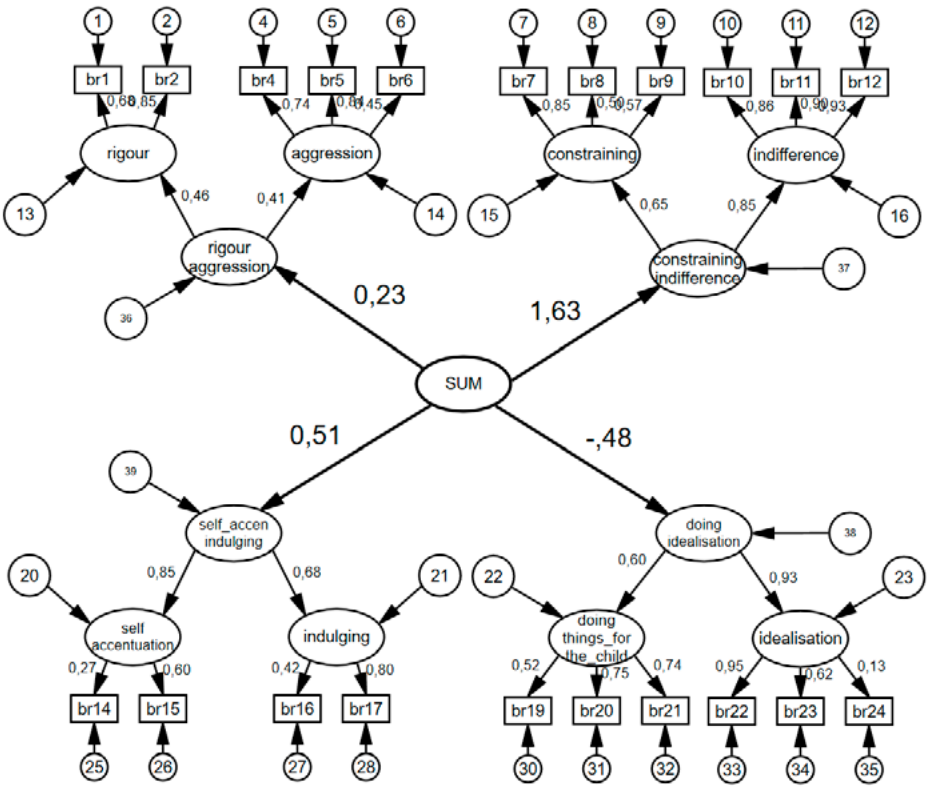


Figure 14.19. Confirmatory factor analysis of the circular model of parental mistakes, with reduction to a general dimension of mistakes

The model reveals a problem in which the value of the relationship between *constraining* and *indifference* and the cold mistakes exceeds 1 (β = 1.28), as well as the correlation φ = 1.30 between cold mistakes and warm mistakes (Figure 14.18), and the relationship between *constraining* and *indifference* and all parental mistakes (β = 1.63) (Figure 14.19). Jöreskog—the creator of structural equation models—states *"A common misunderstanding is that the coefficients in the completely standardized solution must be smaller than one in magnitude and if they are not, something must be wrong. However, this need not be so. […]. By a standardized coefficient I mean any estimated coefficient in a measurement or structural relationship in the completely standardized solution. […]. The misunderstanding probably stems from classical exploratory factor analysis where factor loadings are correlations if a correlation matrix is analyzed and the factors are standardized and uncorrelated (orthogonal). However, if the factors are correlated (oblique), the factor loadings are regression coefficients and not correlations and as such they can be larger than one in magnitude. This can indeed happen also for any factor loading or structural coefficient in any LISREL model. […]. Just remember that a standardized coefficient of 1.04, 1.40, or even 2.80 does*

not necessarily imply that something is wrong, although, as will be seen, it might suggest that there is a high degree of multicollinearity in the data".(Jöreskog, 1999).

To verify the circular model of parental mistakes, a third method was used: the grade algorithm. Table 14.12 presents the original correlation matrix between all mistakes.

Table 14.12. Correlation matrix for parental mistakes

	ryg	agr	ham	oboj	eksp	uleg	zast	ideal
ryg	1.0	0.3413	0.2868	0.0623	0.1723	-0.2546	0.0095	0.1821
agr	0.3413	1.0	0.5355	0.39	0.4349	-0.0258	-0.1404	0.012
ham	0.2868	0.5355	1.0	0.4909	0.4749	0.0727	-0.0954	0.0896
oboj	0.0623	0.39	0.4909	1.0	0.3174	0.2141	-0.3738	-0.4197
eksp	0.1723	0.4349	0.4749	0.3174	1.0	0.0968	0.0861	0.0932
uleg	-0.2546	-0.0258	0.0727	0.2141	0.0968	1.0	0.0769	-0.0587
zast	0.0095	-0.1404	-0.0954	-0.3738	0.0861	0.0769	1.0	0.4201
ideal	0.1821	0.012	0.0896	-0.4197	0.0932	-0.0587	0.4201	1.0

Table 14.13 presents the same correlation matrix, rescaled by a value of one.

Table 14.13. Correlation matrix for parental mistakes rescaled by a value of one

	ryg	agr	ham	oboj	eksp	uleg	zast	ideal
ryg	2.0	1.3413	1.2868	1.0623	1.1723	0.7453	1.0095	1.1821
agr	1.3413	2.0	1.5355	1.3901	1.4349	0.9741	0.8505	1.012
ham	1.2868	1.5355	2.0	1.491	1.4749	1.0727	0.9045	1.0897
oboj	1.0623	1.3901	1.491	2.0	1.3174	1.2141	0.6261	0.5802
eksp	1.1723	1.4349	1.4749	1.3174	2.0	1.0968	1.0861	1.0932
uleg	0.7453	0.9741	1.0727	1.2141	1.0968	2.0	0.0769	0.9412
zast	1.0095	0.8505	0.9045	0.6261	1.0861	0.0769	2.0	1.4201
ideal	1.1821	1.012	1.0897	0.5802	1.0932	0.9412	1.4201	2.0

Table 14.13 was analysed using the grade algorithm, which was applied to perform the relevant calculations. The results of the algorithm, presented as an overrepresentation map, are shown in Figure 14.20.

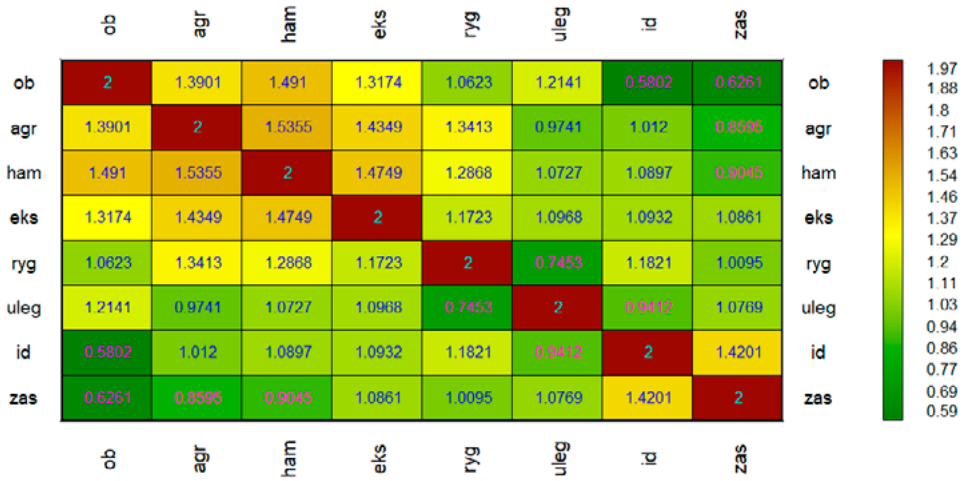


Figure 14.20. Overrepresentation map for parental mistakes

The colour pattern of the map and the absence of the expected circular shape (cf. Figure 14.9) already indicate that the model does not resemble a circular one. For example, the diagonal in the bottom right corner of the overrepresentation map (Figure 14.20) shows that *doing things for the child* displays the strongest correlation with *idealisation*, reaching a correlation coefficient of 0.4201 (Table 14.12). These two variables form a strong overrepresentation, which is consistent with the theoretical assumptions of the parental mistakes model. However, it is worth noting that the mistake of *idealisation* is not related to the other mistakes in the model in the way that was theoretically expected.

Table 14.13 presents both the expected and the obtained correlations for each pair of variables in the matrix, allowing for an evaluation of how closely the model aligns with theoretical assumptions.

Table 14.13. Expected values and observed correlations

Expected values (above) and empirically obtained values (in parentheses) for variable pairs representing parental mistakes in the circular model.

	Rigour	**Aggression**	**Constraining**	**Indifference**	**Self-accentuation**	**Indulgence**	**Doing things for the child**	**Idealisation**
Rigour	1.000 *(1.000)*	0.707 *(0.341)*	0.000 *(0.287)*	−0.707 *(0.062)*	−1.000 *(0.172)*	−0.707 *(−0.255)*	0.000 *(0.009)*	0.707 *(0.182)*
Aggression	0.707 *(0.341)*	1.000 *(1.000)*	0.707 *(0.535)*	0.000 *(0.390)*	−0.707 *(0.435)*	−1.000 *(−0.026)*	−0.707 *(−0.140)*	0.000 *(0.012)*
Constraining	0.000 *(0.287)*	0.707 *(0.535)*	1.000 *(1.000)*	0.707 *(0.491)*	0.000 *(0.475)*	−0.707 *(0.073)*	−1.000 *(−0.095)*	−0.707 *(0.090)*
Indifference	−0.707 *(0.062)*	0.000 *(0.390)*	0.707 *(0.491)*	1.000 *(1.000)*	0.707 *(0.317)*	0.000 *(0.214)*	−0.707 *(−0.374)*	−1.000 *(−0.420)*
Self-accentuation	−1.000 *(0.172)*	−0.707 *(0.435)*	0.000 *(0.475)*	0.707 *(0.317)*	1.000 *(1.000)*	0.707 *(0.097)*	0.000 *(0.086)*	−0.707 *(0.093)*
Indulgence	−0.707 *(−0.255)*	−1.000 *(−0.026)*	−0.707 *(0.073)*	0.000 *(0.214)*	0.707 *(0.097)*	1.000 *(1.000)*	0.707 *(0.077)*	0.000 *(−0.059)*
Doing things for the child	0.000 *(0.009)*	−0.707 *(−0.140)*	−1.000 *(−0.095)*	−0.707 *(−0.374)*	0.000 *(0.086)*	0.707 *(0.077)*	1.000 *(1.000)*	0.707 *(0.420)*
Idealisation	0.707 *(0.182)*	0.000 *(0.012)*	−0.707 *(0.090)*	−1.000 *(−0.420)*	−0.707 *(0.093)*	0.000 *(−0.059)*	0.707 *(0.420)*	1.000 *(1.000)*

It is evident that only a few relationships are fully consistent with the theoretical model—that is, they meet both the expected direction (sign) and the strength of the correlation. Most variable pairs demonstrate only partial consistency (e.g. correct sign but weakened strength), or even significant discrepancies, such as reversed direction or near-zero values where strong associations were expected.

The use of the grade algorithm highlights key shortcomings in the model. The clearly marked red and orange areas on the overrepresentation map indicate significant correlations between specific groups of variables. Located on opposite ends of the map are only two mistakes: *doing things for the child* and *idealisation*. Multidimensional scaling does not depict these relationships as clearly, which underscores the advantage of the grade algorithm and the overrepresentation map method in visualising such differences. The map shows that *doing things for the child* and *idealisation* form a distinct cluster that clearly diverges from the remaining parental mistakes.

While the layout of points in the multidimensional scaling might suggest alignment with the theoretical model, CFA analysis reveals that the structure of *doing things for the child* and *idealisation* breaks away from the other mistakes. Moreover, the overrepresentation map unambiguously shows that these two mistakes are the only ones that correlate negatively with all other parental mistakes—something not evident in the other analyses.

The grade algorithm also provides additional insights not visible in either the CFA models or the multidimensional scaling. For instance, while multidimensional scaling showed that *constraining* the child's activity correlates with *aggression*, the grade

algorithm revealed that *constraining* is also significantly associated with *self-accentuation* by the parent and with *indifference*. These relationships had not been previously detected, emphasising the added value of the grade algorithm in the analysis of parental mistakes.

The grade algorithm also yields information about underrepresentation, which constitutes an important supplement to the model analysis. According to the assumptions of the circular model, variables located on opposite sides of the circle should not be correlated, and those positioned at a 90-degree angle should exhibit no correlation. Neither structural models nor multidimensional scaling guarantee that these relationships will be accounted for, but the overrepresentation map allows for their verification. An example of theoretical consistency is the lack of association between *rigour* and *doing things for the child*—two mistakes positioned at a 90-degree angle. Similarly, *rigour* does not correlate with *indulgence*, which also confirms the model's assumptions, even though in the theoretical layout these variables form a wider angle, suggesting a potential negative correlation.

However, the results obtained from multidimensional scaling turn out to be inconsistent with the findings of the grade algorithm. In light of these results, the assumption that the model under study is fully circular appears highly questionable. Multidimensional scaling only allows for the presumption that the model is *circle-like* in character.

In summary, the application of the grade algorithm and hierarchical confirmatory factor analysis provided valuable additional insights into the relationships between variables in the circular model of parental mistakes—insights that would not have been visible in an analysis based solely on multidimensional scaling. While the scaling suggests some alignment with the assumptions, the grade algorithm reveals significant deviations from the theoretical model, offering a more comprehensive picture of the interrelations.

This indicates that the evaluation of circular models should not rely solely on a single method such as multidimensional scaling. It is essential to combine different analytical techniques—such as hierarchical confirmatory factor analysis and the grade algorithm—which together enable a more accurate assessment of model fit to theoretical assumptions, as well as the identification of key relationships and limitations. When used in isolation, multidimensional scaling may prove insufficient for a comprehensive evaluation of the model's relational structure.

In the next part of the book, an additional method for validating circular models will be presented—one based on the analysis of three-dimensional space. This method may prove especially useful in the case of complex models, in which variables are more densely distributed, or when the results of traditional methods are inconsistent—for instance, in models that appear well-fitted in one measurement but invalid in another. Such a situation may apply to the model of parental mistakes, where fluctuations in the results may suggest that a three-dimensional structure underlies the two-dimensional projections. In such cases, the use of a 3D approach may allow for a more complete capture of the model's dynamics and a better understanding of its structure (Szymańska, 2025a, 2025d).

CHAPTER 15

Clustering Algorithms in Cluster Analysis

Clustering algorithms are key tools in data exploration, used to group datasets based on similarities between objects. In contrast to classification algorithms, which require predefined classes, clustering algorithms operate on data in an unsupervised manner, aiming to identify natural groupings (clusters) within the data. These techniques are extremely useful in many fields, including market analysis, biology, medicine, finance, and the social sciences.

The primary goal of clustering algorithms is to divide a dataset into several clusters such that the objects within the same cluster are more similar to one another than to objects in other clusters (Szymańska, 2017d). This process enables a better understanding of data structure, the discovery of patterns, and the identification of anomalies. In this chapter, we focus on the two most commonly used clustering algorithms: *k*-means and EM (Expectation-Maximization)**.** Each of these algorithms has its own specific characteristics, advantages, and limitations, making their proper selection critical to achieving reliable and meaningful results for a given analytical task.

Clustering algorithms play a crucial role in data analysis because they allow for complexity reduction by grouping objects into clusters. This simplifies the analysis of large datasets and enables more comprehensible interpretation and visualisation of the results. Through clustering, it becomes possible to uncover hidden patterns and relationships in the data, potentially leading to new discoveries and a deeper understanding of the studied problem.

One of the most frequently used clustering algorithms is the *k*-means algorithm. This algorithm operates by iteratively assigning each data point to the nearest cluster centre (centroid) and then updating the positions of the cluster centres based on their current members (Szymańska, 2017d). This process is repeated until convergence is

reached—that is, until the assignments of points to clusters no longer change. *K*-means is valued for its simplicity and efficiency, although it can be sensitive to the choice of initial centroids and the number of clusters.

Another widely used clustering algorithm is the EM algorithm (Expectation-Maximization) (Nisbet et al., 2009). EM is a more advanced technique that estimates the parameters of probability distributions for each cluster and iteratively assigns data points based on maximising the expected value of the likelihood function. It is particularly useful in cases where the data originates from a mixture of different distributions.

Clustering is also used in many advanced techniques of statistical modeling and machine learning. Application examples include customer segmentation, image classification, anomaly detection in financial data, biological analyses such as grouping genes with similar functions, and social research, where behavioural patterns are analysed in large survey datasets.

Thanks to their ability to reveal hidden structures in data, clustering algorithms form the foundation of many modern analytical and exploratory applications. In the following sections of this chapter, we will discuss in detail the techniques and practical applications of the *k*-means and EM algorithms in cluster analysis, which will facilitate a deeper understanding of their operation and potential uses.

15.1. Cluster Analysis Using the k-means Algorithm: Techniques and Practical Examples

The *k*-means algorithm is one of the most popular and widely applied clustering algorithms. Its popularity stems from its simplicity of implementation and operational efficiency (Nisbet et al., 2009).

The *k*-means algorithm operates in several steps. Initially, the number of clusters *k* is selected, and initial centroids (cluster centers) are randomly placed. Then, each data point is assigned to the nearest centroid, thus forming clusters. The next step involves updating the centroids, where new centroids are calculated as the mean value of the points assigned to each cluster. These steps of assigning points to clusters and updating centroids are repeated iteratively until stability is achieved—when the point assignments to clusters no longer change (Szymańska, 2017d).

The *k*-means algorithm seeks to minimize the sum of squared distances between the data points and their respective centroids, which leads to compact and clearly defined cluster divisions. Despite its simplicity, the *k*-means algorithm has certain limitations, such as sensitivity to the selection of initial centroids and the number of clusters, as well as a tendency to form clusters of similar size and shape.

Formally, the *k*-means algorithm can be described as follows. At the outset, the number of clusters *k* is selected, and *k* points from the data space are randomly chosen as initial centroids $\mu_1, \mu_2, \ldots, \mu_k$.

In the next step, each data point x_i is assigned to the cluster C_j whose centroid μ_j is the closest according to a distance metric (most commonly Euclidean distance) (Szymańska, 2017d). Formally, the point x_i is assigned to cluster C_j if:

$$C_j = \left\{ x_i : \|x_i - \mu_j\|^2 \leq \|x_i - \mu_l\|^2 \forall l \in \{1, 2, \ldots, k\} \right\}$$

This means that the point is assigned to the cluster whose centroid lies closest to it in terms of Euclidean distance—that is, the distance between the point and that cluster's centroid is the smallest among all possible clusters.

After all points have been assigned to clusters, the new position of each centroid is calculated as the arithmetic mean of the points assigned to the given cluster. The new centroid μ_j is computed as:

$$\mu_j = \frac{1}{C_j} \sum_{x_i \in C_j} x_i$$

where $|C_j|$ denotes the number of points in cluster C_j.

This means that the new centroid is determined as the center of gravity of all points assigned to that cluster—that is, the point located in the "average" position relative to those data. As a result, the centroid "shifts" toward the actual distribution of data within the cluster, allowing for a better reflection of its structure. This process is repeated iteratively until stabilization—i.e., when the centroids no longer change significantly.

The steps of assigning points to clusters and updating centroids are repeated until stability is reached, either when the assignments of points to clusters stop changing or the changes become negligible. The *k*-means algorithm aims to minimize the objective function known as the within-cluster sum of squares (WCSS):

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

The goal is to find such centroids $\mu_1, \mu_2, \ldots, \mu_k$ that minimize J.

In other words, the objective function measures how far the data points are from their respective centroids. Cluster analysis using the *k*-means algorithm enables the discovery of natural groupings within data, which can be extremely useful in various applications. Typical steps in performing cluster analysis with the *k*-means algorithm include data preparation, selection of the number of clusters, execution of the *k*-means algorithm, visualization of results, and interpretation of the clusters. In the first step, data are collected and prepared, including feature standardization to ensure that all features are on comparable scales (Szymańska, 2017c).

Next, the optimal number of clusters *k* is determined, which can be achieved using the elbow method. This involves analysing the within-cluster sum of squares (WCSS) as a function of the number of clusters (Warchalska-Troll & Warchalski,

2022). The WCSS is plotted for different values of *k*, and a point is sought where further increases in the number of clusters yield diminishing improvements—this point indicates the optimal number of clusters. Once the number of clusters has been established, the *k*-means algorithm is applied to the data, resulting in a division into *k* clusters.

Cluster visualization in the feature space is essential for assessing the quality of the division and understanding the data structure. Dimensionality reduction techniques (e.g., PCA, t-SNE) are often employed to improve visualization in 2D or 3D. Finally, the characteristic features of individual clusters are analysed to understand their meaning and potential applications.

The *k*-means algorithm is valued for its speed and scalability, which makes it well-suited for the analysis of large datasets. In practice, the algorithm is frequently used in customer segmentation, image classification, behavioural pattern analysis, and anomaly detection in financial data. Despite certain limitations, the appropriate application of the *k*-means algorithm can lead to valuable discoveries and improved understanding of the analysed data.

The Statistica Data Miner package enables cluster analysis using the *k*-means method with clustering algorithms. This procedure has been described, among others, by Miner and his team as well as by Szymańska (Nisbet et al., 2009; Szymańska, 2017d). Cluster analysis constitutes the inverse of analysis of variance. Objects are classified into clusters based on the principle of maximizing between-group variance and minimizing within-group variance (Szymańska, 2017d). This means that in practice, the algorithms classify objects so that they are as similar to each other as possible within clusters while aiming to maximize the differences between clusters.

This analysis constitutes the inverse of analysis of variance in the sense that analysis of variance using the least squares method tests whether the differences between groups are statistically significant in a given dataset (Aranowska & Rytel, 2010). In contrast, cluster analysis aims to classify objects according to the rule of maximizing internal cohesion within clusters and maximizing differences between them (Szymańska, 2017d).

Classification algorithms analyse the cluster values of elements across all variables simultaneously. This means that objects are classified based on the principle of maximizing between-group variance and minimizing within-group variance while taking all analysed variables into account. The number of clusters obtained may be determined by the researcher, as the procedure allows them to specify how many clusters should be generated. This enables the determination of such a number of clusters that achieves maximum between-group variance and minimum within-group variance (Szymańska, 2017d).

Although limiting this number or imposing it by the researcher may result in successfully identifying a predetermined number of clusters, such a solution may not be optimal. Therefore, decisions to limit or impose the number of clusters should be carefully considered. The V-Cross Validation procedure, which automatically generates the number of clusters, appears to be a more optimal approach than imposing the number of clusters manually.

It is also not surprising that changes in the number of analysed variables may affect the number of clusters. This is an expected phenomenon. It is not possible to predict precisely how many clusters the algorithm will ultimately generate, as this depends on the specific characteristics of the dataset under analysis.

Analysing the number of clusters is an extremely important aspect of interpreting the results obtained through the clustering procedure. In a sufficiently large random sample, it is possible to attempt to replicate the number of combinations, configurations of feature levels, or profiles characteristic of a given population. This way, one can verify, for example, how many and what kinds of variable configurations occur in the population—which of these configurations are most typical, and which are less frequent.

However, in order for such verification to be possible, in addition to the discussed clustering method, a procedure for determining the fit of model curves to empirical ones should also be applied. This procedure will be described in Chapter 16.

In addition to the number of identified clusters, the distances between clusters and the related shape of profiles (curves) are also subject to interpretation. The method of interpretation will be further explained later in this chapter using a specific example. At this point, it is necessary to return to the formal aspects of the discussed cluster analysis.

The result of cluster analysis is presented using a plot that displays curves representing the clusters in a coordinate system. These curves take the form of profiles. On the *y*-axis, the procedure marks the mean standardized values, calculated using the following formula:

$$\overline{x} = \frac{\overline{x_k} - x_{min}}{x_{max} - x_{min}} \tag{15.1}$$

Where:
$\overline{x_k}$ – the mean of a given group for a particular variable
x_{min} – the minimum score that a respondent could achieve on the given scale
x_{max} – the maximum score that a respondent could achieve on the given scale
$x_{max} - x_{min}$ – the range of scores on the scale

The standardized mean takes on values from 0 to 1, where 0 indicates that none of the individuals in the sample obtained any points for the given variable, and 1 indicates that all individuals in the cluster achieved the maximum score on that variable. Scores from 0 to 0.4 are interpreted as low, from 0.4 to 0.7 as moderate, from 0.7 to 0.9 as high, and from 0.9 to 1 as very high (Szymańska, 2017d, 2019). Presenting the results of cluster analysis in the form of profiles allows for an immediate comparison of which group achieved higher and which achieved lower scores, as well as what those scores were on the scale.

Cluster interpretation is based on the analysis of standardized mean values. In addition to the profile curves representing the clusters, the procedure also provides a table displaying the mean values for each variable and the number of cases classified into each cluster. This enables the identification of which profiles

occur more frequently and which are rarer. In other words, it becomes possible to determine whether any profile is more typical than others.

Differences between clusters are tested using analysis of variance (ANOVA), which makes it possible to determine whether the differences between clusters for each variable are statistically significant. As will be demonstrated in the example, using the information on between-group and within-group sums of squares, the researcher can independently calculate the effect size for the variables within the cluster using effect size measures such as eta squared (η^2) (Szymańska, 2019).

Example of Cluster Analysis Using the k-means Algorithm

To illustrate the application of the *k*-means algorithm in cluster analysis, we present an example based on children's temperamental traits measured using the DOTS-R questionnaire. The data for this analysis were drawn from the sample described in Appendix A. The DOTS-R (Dimensions of Temperament Survey–Revised) is a tool used to assess various aspects of children's temperament, such as activity, rhythmicity, adaptability, and mood (Śliwińska, Zawadzki, &Strelau, 1995).

The analysis was conducted using the *k*-means algorithm implemented through data mining tools in the STATISTICA software. The aim of the analysis was to identify natural groups (clusters) of children based on their temperamental traits. The analytical process included several key steps: initialization, assignment of points to clusters, updating of centroids, and iteration of these steps until stability was achieved.

Figure 15.1 presents the results of the cluster analysis, displaying a graphical representation of the clusters in a coordinate system. Each curve represents a cluster profile, and the *y*-axis values indicate the standardized mean values of temperamental traits for each cluster. This allows for an easy comparison of which groups of children scored higher or lower in particular aspects of temperament.

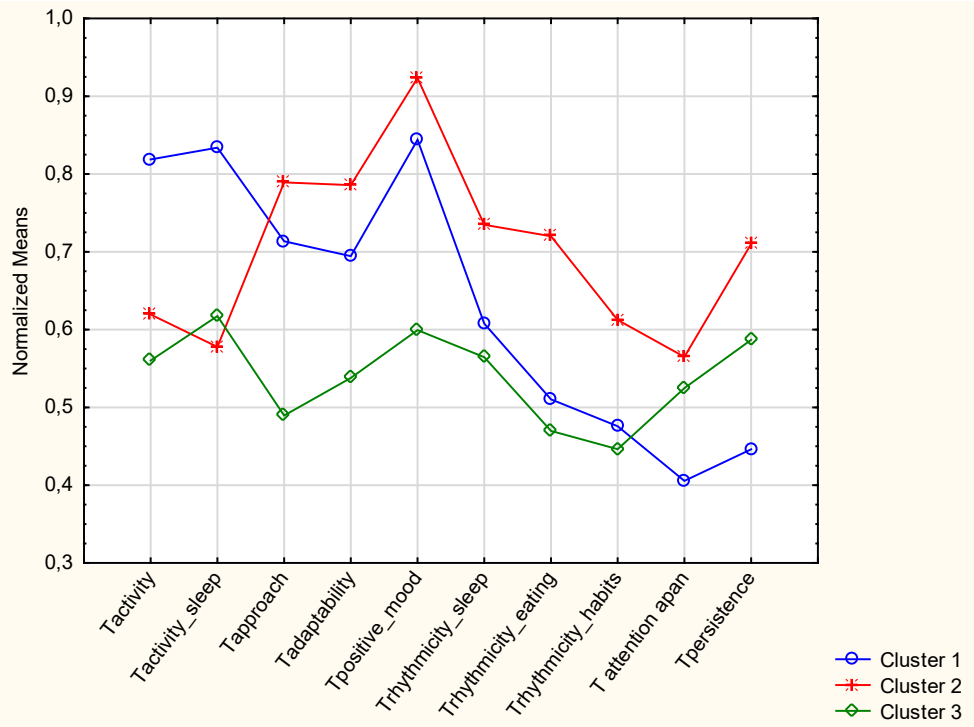


Figure 15.1. Graph of standardized mean variable values for three clusters (k-means)

Cluster 1 includes 33.08% of the sample (133 cases) (Table 15.1) and is characterized by high and moderate values across various variables, suggesting a diverse temperamental profile of the children. In the graph, Cluster 1 is marked with circles.

Table 15.1. Mean variable values for three clusters (*k*-means)

Variable	Cluster 1	Cluster 2	Cluster 3
Tactivity	24.1954887	20.0083333	18.7651007
Tactivity_sleep	14.0075188	10.9333333	11.4093960
Tapproach	21.9849624	23.5750000	17.2818792
Tadaptability	15.7218045	17.0000000	13.5302013
Tpositive_mood	25.5037594	26.7750000	21.5906040
Trhythmicity_sleep	16.9398496	19.2250000	16.1610738
Trhythmicity_eating	12.6541353	15.8000000	12.0469799
Trhythmicity_habits	12.6616541	14.5750000	12.2483221
T attention apan	11.0827068	13.4750000	12.8724832
Tpersistence	7.01503759	9.39166667	8.28187919
Number of cases	133	120	149
Percentage (%)	33.08%	29.85%	37.06%

Children in Cluster 1 exhibit very high levels of physical activity, with a normalized mean value of 0.82. This indicates that they are highly energetic and physically active, which may suggest a need for increased physical engagement and that they may be more demanding in terms of managing their energy levels. Sleep activity is also high, at 0.83, meaning that these children remain dynamic even in a resting state.

In terms of adaptability, children in this group show a moderate ability to adjust to new situations and people, with a value of 0.72. This suggests that they generally cope well but may require some support in more demanding contexts. Flexibility values are also high, at 0.69, indicating that these children possess a strong capacity to adapt to changes and new conditions.

The very positive mood observed in children in Cluster 1, with a high value of 0.85, suggests that they are typically content and happy, which supports peer relationships and general life satisfaction. This is one of the highest values in this group and highlights their positive attitude.

Regarding sleep regularity, the value is 0.6 – a moderate result suggesting stable, though not perfectly regular, sleep patterns. Eating regularity stands at 0.51, and general habit regularity at 0.47 – also moderate values, indicating relatively stable but not fully consistent daily routines. Such routines may contribute to a sense of safety and stability in children.

Unfortunately, attention span, with a value of 0.4, and persistence, with a value of 0.45, are the lowest among all clusters. This indicates that children in this group may experience the greatest difficulty maintaining focus on tasks and may lose interest more quickly in activities that require sustained effort. This is significant, as it may impact both academic performance and daily functioning.

Cluster 2 includes 29.85% of the sample (120 cases) and is characterized by high and moderate values across various variables, presenting a picture of children who are moderately active, adaptable, persistent, and attentive. In the chart, children in Cluster 2 are marked with stars.

Children in Cluster 2 show a moderate level of physical activity, with a normalized mean value of 0.62. This indicates that they are energetic, but not overly so. Sleep activity is slightly lower, at 0.58, which means these children exhibit stable activity patterns even while at rest.

In terms of adaptability, children in this group demonstrate a high capacity to adjust to new situations and people, with a value of 0.8. This suggests that they easily establish contact and quickly adapt to new conditions. Flexibility values are also high, at 0.78, indicating that these children are capable of accommodating change.

The very positive mood of children in Cluster 2, as indicated by the very high value of 0.93, suggests that they are generally content and happy, which supports their peer relationships and overall life satisfaction. This is the highest value in this group, which emphasizes their positive attitude.

With regard to sleep regularity, the value is 0.73 – this is a moderate result that suggests stable, although not perfectly regular, sleep patterns. Eating regularity is 0.72, and habit regularity is 0.61 – these are also moderate values, indicating stable

but not perfectly consistent routines. Stable routines may contribute to a sense of safety and stability in children.

Attention span, with a value of 0.57, and persistence, with a value of 0.71, are the highest among all groups. This indicates that children in this group, although they may have some difficulties maintaining focus on tasks, show relatively high persistence in tasks requiring prolonged effort compared to other groups.

Cluster 3 includes 37.06% of the respondents (149 cases) and is characterized mainly by low and moderate values across various variables, indicating children with moderate levels of activity and persistence. In the graph, children in Cluster 3 are marked with squares.

Children in Cluster 3 show moderate physical activity, with a normalized mean value of 0.56. This indicates that they are fairly active, but not overly so. Sleep activity is slightly higher, at 0.62, which means that these children have moderately stable patterns of activity during rest.

In terms of adaptability, children in this group have a moderate ability to adjust to new situations and people, with a value of 0.49. This suggests that these children may have difficulty adapting to new conditions and may need additional support. Flexibility values are also moderate, at 0.54, indicating that these children have a limited ability to adapt to changes.

The good mood of children in Cluster 3, as indicated by the moderate value of 0.6, suggests that they are generally content, but may be prone to mood fluctuations. This is a moderate value, which highlights their overall, though not always stable, positivity.

With regard to sleep regularity, the value is 0.57, which is moderate and suggests stable, though not perfectly regular, sleep patterns. Eating regularity is 0.47, and habit regularity is 0.44, which are also moderate values, indicating greater variability in daily routines. Lower regularity of habits may affect children's sense of stability and safety.

The ability to maintain attention, with a value of 0.53, and persistence, with a value of 0.59, are moderate. This suggests that children in this group may experience some difficulties in maintaining focus on tasks, yet are capable of demonstrating perseverance in tasks that require prolonged effort.

Table 15.2. Standardized distances between centroids for the three clusters (*k*-means)

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0.000000	0.547462	0.535402
Cluster 2	0.547462	0.000000	0.630328
Cluster 3	0.535402	0.630328	0.000000

The table presenting standardized distances between centroids highlights the differences among the centroids of the respective clusters. The distance between Cluster 1 and Cluster 2 is 0.547, which suggests a moderate difference between these groups. Similarly, the distance between Cluster 1 and Cluster 3 is 0.535, also indicating a moderate

difference. The greatest difference is observed between Cluster 2 and Cluster 3, with a distance of 0.630, which points to a larger divergence between these groups.

Overall, these distances are moderate to large, meaning that the groups differ from one another. In the context of cluster analysis, distances below 0.3 can be considered small, distances ranging from 0.3 to 0.6 are moderate, and distances greater than 0.6 may be regarded as large. These differences are, of course, statistically significant (see Table 15.3).

Table 15.3. ANOVA for quantitative variables after *k*-means clustering

Variable	Between SS	*df*	Within SS	*df*	F	p-value
Tactivity	2218.499	2	6096.687	399	72.5953	0.000000
Tactivity_sleep	718.044	2	2406.486	399	59.5265	0.000000
Tapproach	2952.216	2	5117.456	399	115.0898	0.000000
Tadaptability	837.127	2	2809.821	399	59.4368	0.000000
Tpositive_mood	2014.487	2	3296.200	399	121.9253	0.000000
Trhythmicity_sleep	654.755	2	3306.578	399	39.5042	0.000000
Trhythmicity_eating	1037.551	2	3829.961	399	54.0453	0.000000
Trhythmicity_habits	394.541	2	2186.912	399	35.9918	0.000000
Tattention_span	401.271	2	2010.592	399	39.8159	0.000000
Tpersistence	358.143	2	972.723	399	73.4531	0.000000

The cost sequence chart illustrates the changes in cost values (within-cluster sum of squares, WCSS) depending on the number of clusters (Figure 15.2). The horizontal axis represents the number of clusters, while the vertical axis displays the cost values. The chart shows that as the number of clusters increases, the cost value decreases. This is typical for the *k*-means algorithm, as adding more clusters allows for a better fit of the clusters to the data, which results in a reduction of the within-cluster sum of squares.

With two clusters, the cost is approximately 0.58. This is the highest value, indicating that the data are not yet optimally partitioned. When the number of clusters is three, the cost drops to around 0.55. This represents a significant reduction compared to two clusters, suggesting that three clusters better reflect the structure of the data. At this point, a distinct "bend" in the chart can be observed, which is often referred to as the "elbow". When the number of clusters increases to four, the cost further decreases to approximately 0.53. Although the cost continues to decrease, the reduction is less substantial compared to the shift from two to three clusters.

The chart indicates the presence of an "elbow" at three clusters. The elbow method involves identifying the point at which further increases in the number of clusters yield progressively smaller reductions in cost (Warchalska-Troll & Warchalski, 2022). In this case, the elbow point occurs at three clusters. This means that dividing the data into three clusters is optimal, as it provides a clear cost reduction without significant further improvement when more clusters are added.

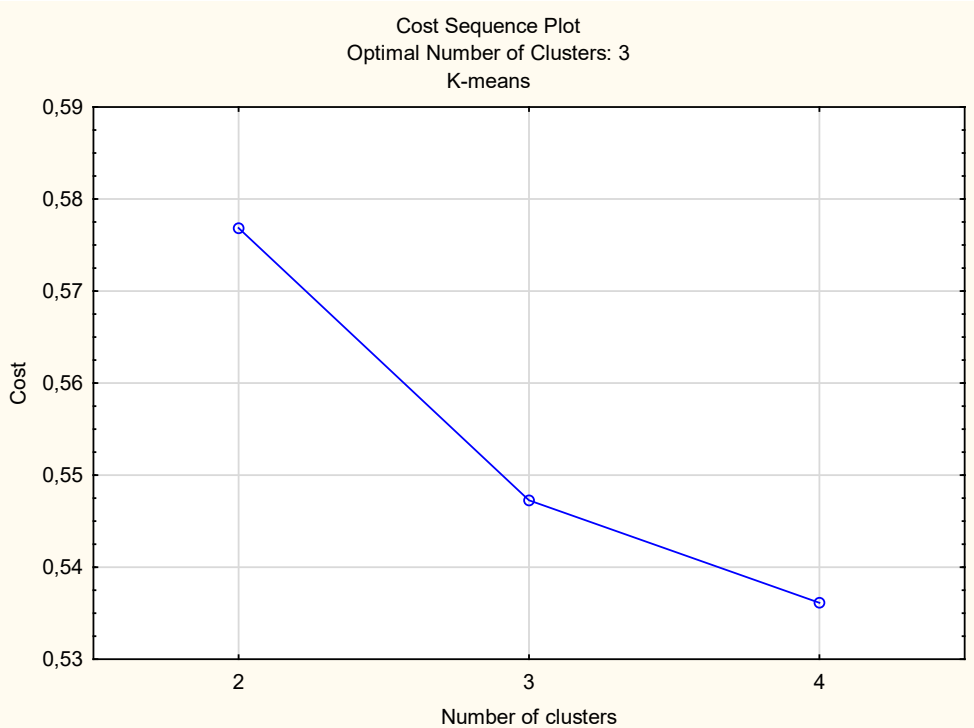


Figure 15.2. Cost sequence chart for varying numbers of clusters (k-means)

The optimal number of clusters for the analyzed data is three. Dividing the data into three clusters offers the best compromise between accurate representation of the data structure and model simplicity. Adding a fourth cluster results in a further decrease in cost; however, the benefit is smaller, suggesting that additional clusters may introduce unnecessary complexity without substantially improving the model.

In practice, selecting the optimal number of clusters is essential for obtaining meaningful and useful results in cluster analysis. In this case, three clusters are sufficient to accurately reflect the structure of the data, which should be considered in the further interpretation and application of the clustering results.

15.2. Cluster Analysis Using the EM Algorithm: Advanced Applications in Statistical Modeling

The Expectation-Maximization (EM) algorithm is an advanced clustering technique that differs from the *k*-means algorithm primarily in that it relies on estimating the parameters of probability distributions for each cluster (Friedman, 2013). It is particularly useful in situations where data are drawn from a mixture of different distributions.

The EM algorithm consists of two main steps: the Expectation Step (E-Step) and the Maximization Step (M-Step). This process is iterative and repeats until convergence (Dempster et al., 1977). Initially, the algorithm assigns random starting values to the parameters of the probability distributions for each cluster. These parameters may include means, variances, and prior probabilities for each cluster.

In the Expectation Step (E-Step), the algorithm computes the probability of each data point belonging to each cluster based on the current parameter estimates (Dempster et al., 1977). For each data point x_i and cluster k, this probability can be expressed as:

$$\gamma_{ik} = \frac{\pi_k N(x_i|\mu_k,\Sigma_k)}{\Sigma\pi_j N(x_i|\mu_j,\Sigma_j)}$$

where γ_{ik} is the probability that point x_i belongs to cluster k, π_k is the prior probability of cluster k, and $N(x_i|\mu_k,\Sigma_k)$is the normal density function for point x_i with parameters μ_k (mean) and Σ_k (covariance). **K** denotes the number of clusters.

In the Maximization Step (M-Step), the algorithm updates the estimates of the distribution parameters for each cluster based on the calculated membership probabilities (Dempster et al., 1977). The parameter updates may be expressed as:

$$\mu_k^{(t+1)} = \frac{\Sigma_{1=1}^N \gamma_{ik} x_i}{\Sigma_{1=1}^N \gamma_{ik}}$$

$$\Sigma_k^{(t+1)} = \frac{\Sigma_{1=1}^N \gamma_{ik}\left(x_i - \mu_k^{(t+1)}\right)\left(x_i - \mu_k^{(t+1)}\right)^T}{\Sigma_{1=1}^N \gamma_{ik}}$$

$$\pi_k^{(t+1)} = \frac{1}{N}\Sigma_{1=1}^N \gamma_{ik}$$

where μ_k is the updated mean for cluster k, Σ_k is the updated covariance matrix for cluster k, π_k is the updated prior probability of cluster k, and N is the number of data points.

The EM algorithm differs from the k-means algorithm in several key aspects. Most importantly, the k-means algorithm assigns points to clusters based on Euclidean distance to centroids, which results in dividing the data into clusters of similar size and shape. In contrast, the EM algorithm assigns points to clusters based on probabilities, estimating the parameters of probability distributions for each cluster. This allows for the modeling of more complex data structures, including clusters of varying shapes and sizes.

The k-means algorithm performs best with data characterized by uniform distribution and similar variances, whereas EM is more flexible and can model data drawn from mixtures of different distributions, making it more suitable for datasets with complex structures. In k-means, each point is assigned to a single cluster in a deterministic way, while in EM, assignment is probabilistic, meaning that each point may belong to more than one cluster with certain probabilities.

The *k*-means algorithm converges quickly but may reach local minima, whereas the EM algorithm may require more iterations to converge, yet it often finds better solutions, especially when the data are complex.

In summary, the EM algorithm is a powerful tool for clustering, especially in the case of data with complex structures and varying distributions. Its probabilistic approach to point assignment and estimation of distribution parameters makes it more flexible than the *k*-means algorithm. Although it is more complex and may require greater computational effort, the EM algorithm offers more accurate and diverse cluster models, which makes it a valuable tool in data exploration and cluster analysis.

Example of Cluster Analysis Using the EM Algorithm

To compare the *k*-means algorithm with the EM (Expectation-Maximization) algorithm, the computations were performed on the same dataset regarding children's temperamental traits using the EM algorithm. This made it possible to compare the results obtained with both algorithms and to assess which approach better models the structure of the data. This comparison allows for identifying differences and similarities in the outcomes, as well as understanding how various clustering algorithms influence cluster analysis and the interpretation of complex datasets.

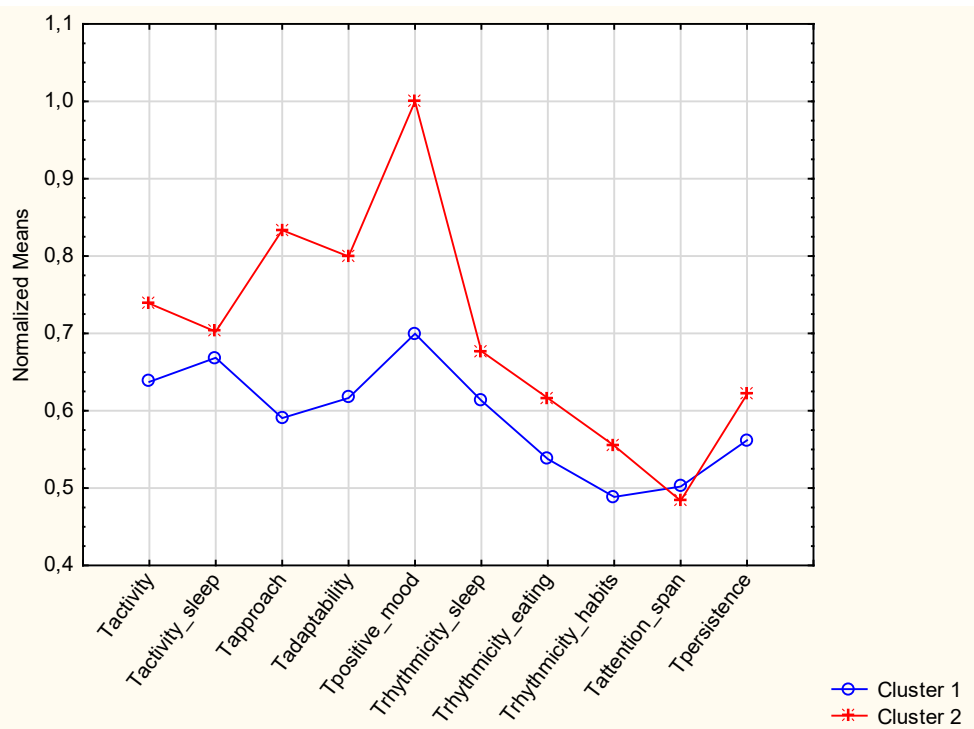


Figure 15.3. Mean Standardized Values for Clusters 1 and 2 According to the EM Algorithm

Figure 15.3 presents the results of cluster analysis conducted using the EM algorithm, which identified two clusters in the dataset. The mean standardized values for each cluster are marked as follows: Cluster 1 – blue circles, Cluster 2 – red stars.

Cluster 1 is characterized by moderate values across most variables. Physical activity has a mean standardized value of 0.65, indicating moderate activity. Activity during sleep is at a similar level of 0.67. Approach toward others scores 0.60, also reflecting a moderate level. Adaptability is 0.61, suggesting that individuals in this cluster are moderately adaptable. Positive mood reaches 0.70, which indicates a moderately high level of satisfaction. Rhythmicity of sleep, eating, and habits are all moderate, at 0.61, 0.54, and 0.48 respectively. Attention span and persistence are also moderate, with values of 0.51 and 0.57 respectively.

Cluster 2 displays higher values in most variables. Physical activity reaches 0.74, indicating a high level of energy. Activity during sleep is also high, at 0.70. Approach toward others is 0.84, suggesting that individuals in this cluster are very sociable. Adaptability reaches 0.80, meaning these individuals are highly flexible and adapt easily to change. Positive mood is at 1.0, indicating the highest level of satisfaction and happiness. Rhythmicity of sleep, eating, and habits are also high, at 0.68, 0.63, and 0.56 respectively. Attention span is at 0.49, and persistence reaches 0.62, which are moderate values.

The ANOVA for quantitative variables confirms the statistical significance of the differences between clusters in most variables (see Table 15.4). The results indicate that differences in variables such as physical activity, approach toward others, adaptability, positive mood, rhythmicity of sleep, eating, and habits are statistically significant. No significant differences were found in variables such as activity during sleep, attention span, and persistence.

Table 15.4. ANOVA for quantitative variables after clustering using the EM algorithm

Variable	Between SS	*df*	Within SS	*df*	F	*p*-value
Tactivity	647.829	1	7667.357	400	33.7968	0.000000
Tactivity_sleep	2.075	1	3122.455	400	0.2658	0.606445
Tapproach	4297.535	1	3772.137	400	455.7136	0.000000
Tadaptability	1282.481	1	2364.467	400	216.9591	0.000000
Tpositive_mood	705.345	1	4605.341	400	61.2633	0.000000
Trhythmicity_sleep	108.613	1	3852.720	400	11.2765	0.000860
Trhythmicity_eating	294.822	1	4572.690	400	25.7898	0.000001
Trhythmicity_habits	89.321	1	2492.132	400	14.3364	0.000176
Tattention_span	3.650	1	2408.213	400	0.6063	0.436636
Tpersistence	4.521	1	1326.345	400	1.3633	0.243658

The EM (Expectation-Maximization) algorithm adopted the following prior probabilities (class weights) for the two clusters in the dataset: for Cluster 1, the probability is 0.742225989, which means that the algorithm initially assumes that

approximately 74.22% of cases belong to this group, while for Cluster 2, the probability is 0.257774011, indicating that approximately 25.78% of cases are assumed to belong to this group.

Prior probabilities in the EM algorithm represent initial estimates of the proportion of data belonging to each cluster. These are the starting weights used by the algorithm to iteratively adjust the model in order to maximize the likelihood of the data. These values influence the initial assignment of cases to clusters and are used in the first phase of the EM algorithm, known as the Expectation Step (E-Step). In this phase, the algorithm computes the expected values of case assignments to clusters, taking into account both the prior probabilities and the observed data.

Next, in the Maximization Step (M-Step), the algorithm updates model parameters such as means and variances for each cluster based on the assignments calculated in the E-Step. In this phase, the prior probabilities may also be updated as the proportions of cases assigned to each cluster. The EM algorithm iterates between the E-Step and the M-Step until convergence is achieved—that is, until the changes in case assignments and model parameters become minimal.

Prior probabilities are therefore crucial for the stability and convergence of the EM algorithm. Proper estimation of these values can accelerate the convergence of the algorithm and improve the quality of clustering, which is essential for obtaining reliable and useful results in data analysis. Cases in Cluster 1 typically have membership probabilities close to one, indicating high classification certainty. The variable values are moderate, which is consistent with the overall nature of this cluster. Similarly, cases in Cluster 2 show high membership probabilities, also indicating high classification certainty. The variable values are higher, reflecting more intense characteristic traits of this cluster.

As a result, for each case the algorithm calculates two probabilities—one for Cluster 1 and one for Cluster 2. The case is then assigned to the cluster with the higher probability. The number of cases classified into Cluster 1 is 176, which constitutes approximately 43.78% of all cases. The number of cases classified into Cluster 2 is 226, or approximately 56.22% of all cases.

The discrepancy between the initial prior probabilities and the actual proportions of cases in each cluster may reflect several important aspects. First and foremost, it may indicate that the initial assumptions were inaccurate. The initial prior probabilities may have been overly optimistic or pessimistic in relation to the true structure of the data. This could result from an improper preliminary estimation of the proportions of data in each cluster. If these assumptions were too far from the actual proportions, the algorithm had to perform more iterations to adjust the model to the real data, which could affect its efficiency.

At the same time, the adaptability of the EM (Expectation-Maximization) algorithm allows it to iteratively adjust its assumptions based on the actual data. The final results indicate that the algorithm was able to improve upon its initial assumptions and find more appropriate case assignments to clusters. A large discrepancy

suggests that the algorithm was effective in adapting, but the initial assumptions were imprecise.

This discrepancy may also suggest that the data have a more complex structure than initially assumed. The initial prior probabilities may not have accurately reflected the true distributions of the data across clusters.

In conclusion, the discrepancy between the initial prior probabilities and the actual proportions of cases in the clusters means that the initial assumptions were inaccurate, but the EM algorithm successfully adjusted to the actual data. The EM algorithm is capable of iteratively refining its assumptions and achieving accurate results, even when the starting assumptions are flawed.

Summary of Cluster Analysis

The comparison between the *k*-means and EM algorithms revealed significant differences in the number of identified clusters and in the characteristics of those clusters. The *k*-means algorithm identified three clusters, while the EM (Expectation-Maximization) algorithm identified two. The variable values in the clusters obtained using the EM algorithm exhibit greater diversity, which can be attributed to the algorithm's probabilistic approach to assigning points to clusters.

Although the *k*-means algorithm is simple and converges quickly, it is sensitive to the initial selection of centroids and the number of clusters. In contrast, the EM algorithm offers greater flexibility, enabling the modeling of data originating from a mixture of different distributions, which makes it more suitable for analyzing complex data structures. The discrepancy between initial prior probabilities and the actual proportions of cases in the clusters in the EM algorithm confirms its effectiveness in adapting to real data, even when the initial assumptions are inaccurate.

Both algorithms have unique advantages and limitations. The *k*-means algorithm is more efficient for analyzing data with simpler structures, while the EM algorithm offers more accurate and diverse cluster models for more complex data.

The choice of the optimal clustering algorithm depends on the nature of the data and the aim of the analysis. The *k*-means algorithm is characterized by its ease of implementation and rapid convergence, which makes it an effective tool for analyzing large datasets where partitioning based on minimizing Euclidean distances to centroids is sufficient. However, its results may depend on the initial choice of centroids and the number of clusters, which can pose challenges in the case of more complex data.

Due to its probabilistic approach, the EM algorithm is more flexible and better suited for modeling data derived from mixtures of different distributions. This allows for more complex cluster structures and more accurate point assignments. Nevertheless, the EM algorithm requires more iterations to converge and greater computational power, which may be a limitation when dealing with very large datasets.

For data with simpler structure, where the number of clusters is easy to estimate, the *k*-means algorithm is an appropriate choice due to its speed and efficiency

(Nathiya et al., 2010). In the case of complex and heterogeneous data, the EM algorithm provides more precise and differentiated cluster models, allowing for more accurate modeling and a better understanding of complex dependencies in the data.

In the context of analyzing children's temperamental traits, where the data may exhibit complex structure, the EM algorithm appears to be more suitable. Its ability to model different probability distributions and its flexibility in assigning points to clusters provide more precise and useful results. Although the *k*-means algorithm is faster and simpler, its limitations in terms of cluster homogeneity and sensitivity to initial centroids may affect the quality of results in the case of complex data.

Therefore, in the analysis of children's temperamental traits, the EM algorithm is recommended as the more appropriate approach, offering more accurate cluster models and better fit to the complex structure of the data.

CHAPTER 16

Application of Clustering Algorithms in Profile Curve Modeling

This chapter discusses the application of clustering algorithms to the modeling of empirical profile curves. Particular attention is given to cluster analysis methods and their role in generating profiles that reflect the intensity of variable values within clusters identified by these algorithms. Such profiles are useful not only for describing phenomena within a population but also for testing theoretical assumptions regarding the existence of specific combinations of traits in the population.

By using standardized means, clustering algorithms present results in the form of graphical profiles that visualize the intensity of variable values within a given cluster. This chapter demonstrates how these profiles may be used for both research and practical purposes.

These profiles allow for a detailed description of phenomena in the population, indicating how particular features are distributed across different segments. In other words, they enable the identification of profiles most characteristic for the studied group. Moreover, cluster analysis constitutes a valuable tool for testing theoretical hypotheses by comparing empirical profiles with theoretical ones. This type of verification is essential in empirical sciences, where it is necessary to confirm that theoretical assumptions are reflected in actual data.

Additionally, the results of cluster analysis can be used to complement and support the results of structural equation modeling (SEM). They enable the creation of more detailed descriptions of the characteristics of the studied individuals in relation to the variables analyzed in the SEM model.

The chapter also presents an original proposal for using clustering algorithms to construct empirical profile curves based on real population data. The procedure

for verifying the consistency of such profiles with theoretical assumptions is discussed. A further proposal involves combining cluster analysis methods—such as *k*-means and EM—with structural equation models, which represents an innovative approach to data analysis.

In summary, the use of clustering algorithms in modeling profile curves opens up new possibilities for both scientific research and practical applications. These methods not only facilitate the understanding of complex phenomena but also support the development of scientific theories based on empirical data.

16.1. Fitting Model Curves to Empirical Curves: Methods and Use Cases

The issue of fitting theoretical models to empirical data is a topic with broad application in psychological sciences, as reflected in the growing number of studies in this area. Theoretical models most often take the form of systems that represent the structure of psychological characteristics or the relationships between various traits (Jonkisz, 1998; Szymańska, 2016b). Confirmatory factor analysis (CFA) is used to determine the structure of psychological traits, while structural equation models (SEM) are applied to identify relationships among traits (Aranowska, 1996; Bartholomew et al., 2008; Hair et al., 2006; Szymańska, 2016b, 2016a).

However, the issue of model fitting is rarely addressed in the context of models expressed as curves representing multidimensional psychological characteristics, such as personality profiles or coping styles. This chapter presents a method for assessing the fit between theoretically defined curves—so-called model curves—and curves defined at the empirical level.

Reconstructing Model Curves

In psychology, the plotting of individual profiles is used to diagnose traits that are multidimensional and uncorrelated. This means that the results from individual scales do not add up to one overall score for the respondent (Anastasi & Urbina, 1999). In such cases, diagnosis must take into account the results for each scale separately, while simultaneously being based on all the scores obtained.

For multidimensional and uncorrelated scales, plotting profiles serves an important diagnostic function. It allows for a visual representation of results, making it possible to observe both strongly and weakly developed dimensions. Diagnosis of multidimensional traits is rarely based on the results of individual dimensions. Instead, it usually takes into account specific combinations of traits, as in the case of diagnoses based on the MMPI-2, where, for example, the neurotic or psychotic triad is interpreted (Kucharski & Gomula, 1998).

For certain types of diagnosis, the information about an individual's score must reflect their results across multiple dimensions simultaneously. One example

is the theory of parental mistakes by Antonina Gurycka, according to which parental overprotectiveness is a combination of three mistakes: idealizing the child, replacing the child, and yielding to the child (Gurycka, 1990). However, an analysis limited to just these three variables may lead to oversimplifications and misinterpretations. What if a parent who is overprotective also displays high levels of aggression or indifference? Can such a parent still be considered overprotective? To avoid such inconsistencies, it is necessary to take into account the complete profile of the individual, including all available variables. Even if some variables appear to be key, ignoring others may result in diagnostic errors.

In summary, overprotective parents tend to score high in idealizing the child, replacing the child, and yielding to the child, low in aggression and indifference, and average in rigor, constraining the child's activity, and self-promotion. However, only a full analysis of all variables allows for an accurate determination of the extent to which a given profile fits the overprotectiveness schema and whether other variables indicate the need for a revised classification. This perspective is essential in both theoretical and practical contexts. Possible profiles for a group of overprotective parents toward their children are presented in Figure 16.1.

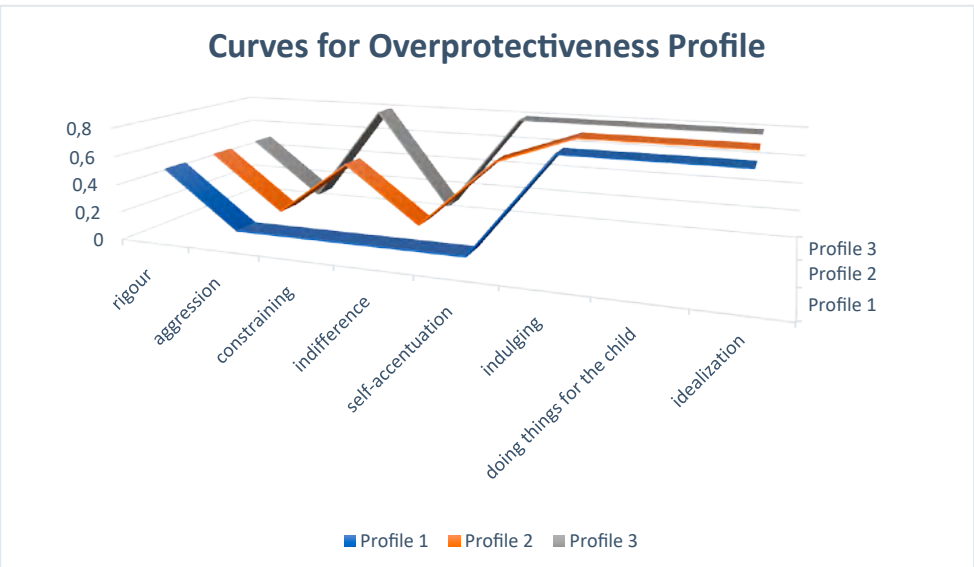


Figure 16.1. Possible model curves for parental mistakes among overprotective parents

As shown, the profile of overprotective individuals indicates that the person displays specific levels across various parental mistakes. In the case of individuals with the trait of overprotectiveness, diagnostically relevant indicators are high scores in idealizing the child, replacing the child, and yielding to the child, and low scores in indifference and aggression. In contrast, the levels of rigor, constraining the child's activity, and self-promotion by the parent are not diagnostic of overprotectiveness. Some overprotective parents may score low on these scales, while others may score

high. Therefore, in the aggregated profile of all overprotective individuals, these variables may take on moderate values, in accordance with the law of regression toward the mean (see Figure 16.1). In practice, however, if theoretical knowledge does not predict any particular outcome for a given trait, it may assume any value but should not be the subject of diagnostic interpretation.

Theoretical knowledge about various psychological traits makes it possible to reconstruct model curves, just as was done for overprotectiveness based on parental mistakes (Figure 16.1). Such reconstruction of model curves for specific traits may have important implications not only for developing theoretical knowledge regarding their co-occurrence with other traits, but also for diagnostic practice.

Plotting trait profiles may lead to a specific theoretical description of those traits. However, the profile must be verified—that is, it must be assessed whether the theoretical assumptions are consistent with empirical data. Model (theoretical) curves should therefore be subjected to empirical verification. If they are found to match empirical data, this means they describe reality accurately and represent valid models. If the theoretical curves do not fit the data, then the assumptions must be revised and reconsidered.

This assumption applies to all types of modeling. Theoretical models are subjected to empirical verification, as in the case of structural equation modeling, where the structural model derived from theory is subsequently tested against data (Bartholomew et al., 2008; Hair et al., 2006; Szymańska, 2016b, 2016a). The same applies to modeling using data mining algorithms. In one part of the dataset, algorithms generate rules—often in the form of a model, such as a decision tree—and in another part, they test the validity of those rules (Elder et al., 2012; Nisbet et al., 2009). The approach proposed here is novel in terms of applying model and empirical curves to test specific assumptions about the co-occurrence of levels in multidimensional variables, though not in terms of the methodology of model fitting. The methodological foundations of this approach are well established in the sciences.

Sometimes a model curve can be constructed based on one of two or more empirical curves (Aranowska, 1989). This occurs when the aim is to assess the degree of fit between two empirical curves. One example is the analysis of parental mistakes. Psychologists working in this area are interested in whether children perceive their parents' mistakes in a similar way as the parents themselves. Gurycka and Wójtowicz conducted research on the intensity of parental mistakes reported by parents. Opinions on these mistakes were collected from both the parents and their children. The results showed that parents and children had significantly different perceptions of parental mistakes. Parents showed almost no diagnostic tendencies in responding to questions about constraining their children's activity, while children identified it as one of the most common mistakes committed by their parents (Wójtowicz, 1989).

A similar issue is addressed in this chapter. Using an example of matching two empirical curves, the method of determining their compatibility is presented. The curves were generated by the *k*-means algorithm, while their convergence was tested using indicators proposed by Elżbieta Aranowska to evaluate curve fit (Aranowska, 1989).

Using Generalized k-means Cluster Analysis to Plot an Empirical Curve

Plotting a model curve based on theoretical knowledge does not present particular challenges. Theoretical curves are derived from assumptions grounded in previous research, scientific theories, or well-established models. In such cases, the researcher has full control over the assumptions and structure of the model, allowing for their precise definition (Szymańska, 2023b). However, this process requires in-depth knowledge of the literature, a clear definition of key variables, and an understanding of their interrelations. The assumptions must be coherent and based on solid theoretical foundations so that the model graph can serve as a valid reference point for future empirical analyses.

In contrast, plotting an empirical curve must meet several key requirements. First, it is necessary to have a sufficiently large and representative sample. Small samples can lead to errors resulting from data randomness and hinder the generation of reliable profiles. Sample representativeness is equally important to ensure that the findings can be generalized to the entire population.

Second, appropriate measurement instruments must be used—tools capable of capturing all relevant variables with precision and reliability. Imperfect diagnostic tools that fail to measure certain variables or do so inadequately may result in an incomplete or distorted representation of the profile.

Third, the data analysis must take into account the full profile rather than focusing only on selected variables. A common mistake in the literature is limiting the analysis to a few characteristic features while overlooking others that may be crucial for interpretation. This means that researchers must include all available data and consider variables not directly measured by the diagnostic tools but still relevant to the model in question.

Fourth, accurate fitting of the empirical curve to the collected data is required. This process involves the application of advanced analytical methods such as cluster analysis. However, maintaining a balance between model simplicity and fidelity to reality can be challenging. Oversimplified models may overlook important nuances, while overly complex models may be difficult to interpret.

Finally, plotting an empirical curve requires careful interpretation of the results. It is essential not only to visualize the data but also to understand their meaning in the context of the research problem. For example, if the profile of an overprotective parent includes high scores in idealizing the child, replacing the child, and yielding to the child—but also elevated scores in aggression—then the individual's classification may need to be reconsidered. Such complexity requires considering the entire profile rather than focusing on only a few selected traits.

In summary, although plotting a model curve based on theoretical knowledge appears to be a relatively straightforward task, the process of constructing an empirical curve demands significantly more effort and attention. The ultimate goal is to create a profile that not only reflects empirical reality but also supports the advancement of scientific theory and practical applications in diagnostics.

Thus, an empirical curve should meet several criteria:

1. It should be based on a large number of individuals to ensure that the resulting shape is reliable.
2. It should take into account the existence of other potential curves that may occur in individuals with certain traits, even if those curves are less frequently represented in the population.
3. It should indicate which curve is more commonly represented and whether statistically significant and substantial differences exist between them.

Reconstructing the empirical curve presents a considerable mathematical challenge. However, thanks to the advancement of clustering algorithms—that is, algorithms that group objects based on their mutual similarity—empirical curves can now be generated with relative ease. To this end, it is proposed to use generalized *k*-means cluster analysis or the Expectation-Maximization (EM) algorithm, both of which are available in STATISTICA software (Elder et al., 2012; Szymańska, 2017d). These methods were described in detail in Sections 16.1 and 16.2 of this book.

An empirical curve generated through cluster analysis can subsequently be used to verify whether it fits the theoretical model curve.

Testing the Fit of a Model Curve to an Empirical Curve

This section presents the procedure for evaluating the fit between a model curve and an empirical curve. It is worth noting that this process serves a similar function to testing the fit of theoretical models to empirical data using structural equation modeling (SEM) (Aranowska, 2005; Bartholomew et al., 2008; Hair et al., 2006; Heck & Thomas, 2009; Szymańska, 2016b). It enables the verification of whether the theoretically constructed model adequately fits the data and is therefore valid.

The method for curve fitting developed by Aranowska, based on fit indices expressed in mathematical formulas, provides a key tool for assessing the alignment of a model curve with an empirical curve (Aranowska, 1989). The method involves determining the closeness of the empirical curve to the model curve. There may be more than one model curve, as different configurations of variable values can reflect a given psychological characteristic. Therefore, several model curves may be derived (as shown in the example of overprotectiveness; see Figure 16.1). It is thus reasonable to define multiple theoretical curves. **The η-Aranowska proximity index** allows for the comparison of an empirical curve with two model curves simultaneously. The index is calculated using Formula 16.1:

$$\eta = \frac{\Sigma(y_i - \tilde{y}_i)^2 - \Sigma\left(y_i - \hat{y}_i\right)^2}{\Sigma\left(\tilde{y}_i - \hat{y}_i\right)^2} \tag{16.1}$$

where:
y_i – value of the comparison system (empirical curve)
$\tilde{y}_i$ – value of the first model system
$\hat{y}_i$ – value of the second model system

The coefficient is therefore calculated as the difference between the sum of squared distances from the empirical curve to the first model curve ($\tilde{y}_i$) and the sum of squared distances to the second model curve ($\hat{y}_i$), divided by the total sum of squared distances between the empirical curve and both model curves.

The interpretation of the ***η*-Aranowska coefficient** is as follows:

- ***η*=0:** The empirical curve is equally distant from both model curves.
- **0<*η*<1:** The empirical curve is more similar to the second model curve ($\hat{y}_i$)
- ***η*=1:** The empirical curve is identical to the second model curve ($\hat{y}_i$)
- **−1<*η*<0:** The empirical curve is more similar to the first model curve ($\tilde{y}_i$)
- ***η*=−1:** The empirical curve is identical to the first model curve ($\tilde{y}_i$)
- ***η*>1:** The empirical curve becomes increasingly similar to the second model curve ($\hat{y}_i$) while significantly differing from the first.
- ***η*<−1:** The empirical curve becomes increasingly similar to the first model curve ($\tilde{y}_i$) while significantly differing from the second.

In practice, the η-Aranowska index can be used to indicate which of the two alternative model curves better describes the empirical data. This analysis is particularly useful in situations requiring the selection of the most appropriate model from several possibilities. The coefficient can also reveal the degree of mismatch between the empirical curve and a given model. However, the resulting values may be ambiguous.

To address this ambiguity, the analysis is complemented by the ζ-Aranowska coefficient, calculated using Formula 16.2:

$$\zeta = \frac{s^2_{\hat{y}_i} + s^2_{y_i} - \frac{1}{2(k-1)} \Sigma \left(y_i - \hat{y}_i\right)^2}{2 \cdot s^2_{\hat{y}_i}} \tag{16.2}$$

where:

$s^2_{\hat{y}_i}$ – variance of the model system
$s^2_{y_i}$ – variance of the empirical system being compared
y_i – value of the empirical system
$\hat{y}_i$ – value of the model system
k – scale multiplicity of y_i

The curves are considered perfectly aligned when ζ=1. This occurs when the variance of the model (theoretical) curve and the variance of the empirical curve are equal, and when half the squared distance between the curves, weighted by degrees of freedom, equals zero.

Therefore, when:

$$s^2_{\hat{y}_i} = s^2_{y_i}$$

$$\frac{1}{2(k-1)} \Sigma \left(y_i - \hat{y}_i\right)^2 = 0$$

When the ζ coefficient takes the value of 0, it indicates that the curves differ from each other only in terms of their variances. In such a case, they may intersect, but they do not converge, or they may be mirror images of each other. The ζ coefficient equals 0 when the sum of the variances of the model (theoretical) curve and the empirical curve is equal to half of the squared distance between the curves, weighted by degrees of freedom. That is, when:

$$s^2_{\hat{y}_i} + s^2_{y_i} = \frac{1}{2(k-1)} \cdot \Sigma\left(y_i - \hat{y}_i\right)^2$$

When $\boldsymbol{\zeta}$ takes values less than 0, this indicates that the curves differ significantly. This occurs when the sum of the variances of the model (theoretical) curve and the empirical curve is less than half of the squared distance between the curves, weighted by degrees of freedom. That is, when:

$$s^2_{\hat{y}_i} + s^2_{y_i} < \frac{1}{2(k-1)} \cdot \Sigma\left(y_i - \hat{y}_i\right)^2$$

The $\boldsymbol{\zeta}$ coefficient takes values greater than 0 when the sum of the variances of the model (theoretical) curve and the empirical curve is greater than half of the squared distance between the curves, weighted by degrees of freedom. That is, when:

$$s^2_{\hat{y}_i} + s^2_{y_i} > \frac{1}{2(k-1)} \cdot \Sigma\left(y_i - \hat{y}_i\right)^2$$

The ζ-Aranowska coefficient is meaningful within the interval $\langle 0, 1\rangle$, as this is the range in which the curves converge.

Examples of Application

Below are two examples of fitting a model curve to an empirical one, based on real psychological research. The first example discusses the fit between a theoretical model curve and an empirical curve, where the model curve was reconstructed at the theoretical level. The second example focuses on the fit between curves when the model curve is also empirical.

First example.
Some psychological theories suggest that stress and difficulties experienced by parents increase the number of parental mistakes they commit (Gurycka, 1990). Previous studies have confirmed that experiencing parental difficulties is associated with constraining the child's activity (Szymańska & Aranowska, 2016). At the theoretical level, two model curves were developed to illustrate the intensity of parental mistakes and the experience of parental difficulties (Figure 21.2). The data used in the analysis come from the sample described in Appendix A.

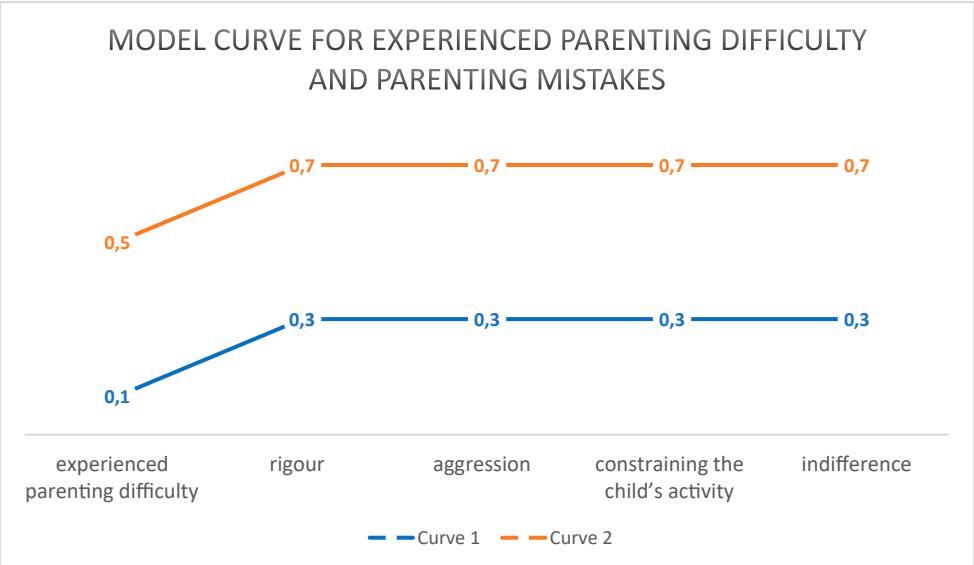


Figure 16.2. Model curves for cold parental mistakes and the experience of parental difficulties

Model curve 1 (denoted by the symbol $\tilde{y}$) takes the following values:
$\tilde{y}_{diff} = 0{,}1$; $\tilde{y}_{rig} = 0{,}3$; $\tilde{y}_{agr} = 0{,}3$; $\tilde{y}_{con} = 0{,}3$; $\tilde{y}_{indif} = 0{,}3$;
Model curve 2 (denoted by the symbol $\dot{y}$) takes the following values:
$\dot{y}_{diff} = 0{,}5$; $\dot{y}_{rig} = 0{,}7$; $\dot{y}_{agr} = 0{,}7$; $\dot{y}_{con} = 0{,}7$; $\dot{y}_{indif} = 0{,}7$

The second model curve reflects a high level of parental difficulties, whereas the first reflects a low level of such difficulties. It can be observed that the expectation of moderate difficulties (a normalized mean in the range of 0.4–0.6) is associated with an elevated level of parental mistakes (a normalized mean above 0.6). In turn, a low experience of difficulties is associated with a low level of parental mistakes. Both curves are parallel.

Subsequently, using Generalized *k*-means cluster analysis, empirical curves were plotted. They are presented in Figure 16.3 and reflect the number of clusters identified in the study sample.

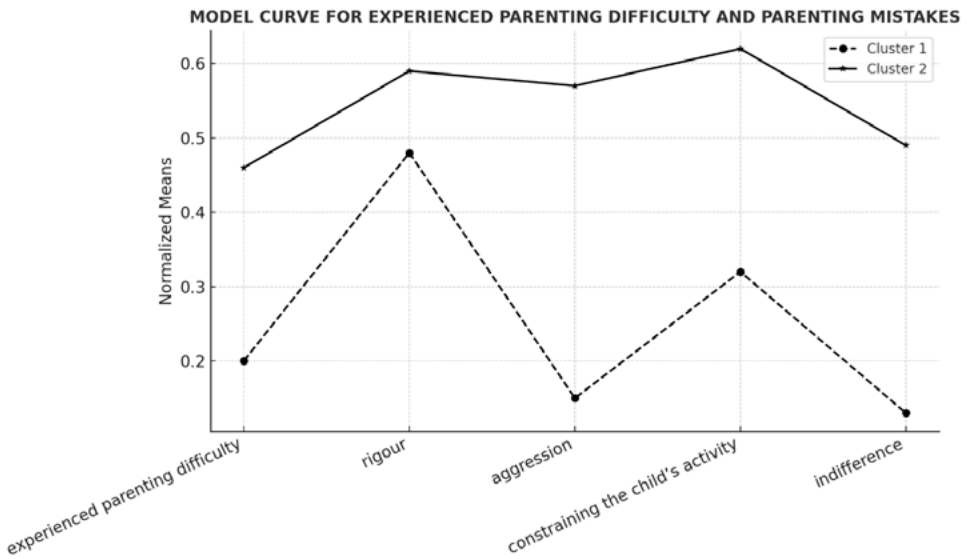


Figure 16.3. Empirical curves for parental mistakes and the experience of parental difficulties

The first cluster represents an empirical curve corresponding to a low level of parental difficulties. It includes 71.9% of the studied group of parents (Table 16.1). At a low level of difficulty, low levels of aggression, constraining the child's activity, and indifference were observed. Parents in this cluster obtained moderate results for rigour (normalized means at the level of 0.5; see Figure 16.3).

Table 16.1 Cluster means, number of cases in clusters, and percentage of cases in the analysis of experienced parental difficulties and parental mistakes

Variable	Cluster 1	Cluster 2
Experienced parental difficulty	11.40909	25.30233
Rigour	15.64545	17.79070
Aggression	5.20000	10.90698
Constraining child's activity	7.53636	11.62791
Indifference	5.69091	10.88372
Number of cases	110	43
Percentage (%)	71.89542	28.10458

The second cluster represents an empirical curve corresponding to a high level of parental mistakes. It includes 28.1% of the studied group of parents. The hypothesis was confirmed that individuals who experience higher parental difficulties also commit more mistakes in the form of rigour, aggression, constraining the child's activity, and indifference (Table 16.1).

There are statistically significant differences between the two clusters (curves) in terms of both parental mistakes and experienced parental difficulties. Effect sizes between the clusters are moderate only for rigour. For the remaining variables, the differences are large (Table 16.2).

Table 16.2 ANOVA results for clusters in the analysis of experienced parental difficulties and parental mistakes

Variable	Between-group variance	*df*	Within-group variance	*df*	*F*	*p*-value	η^2	η^2 Interpretation
Experienced parental difficulty	5967.280	1	15041.66	151	59.9042	< 0.001	0.284	large
Rigour	142.273	1	1898.29	151	11.3172	0.000974	0.070	medium
Aggression	1006.890	1	875.23	151	173.7152	< 0.001	0.535	large
Constraining child's activity	517.540	1	1215.40	151	64.2986	< 0.001	0.298	large
Indifference	833.633	1	1149.91	151	109.4682	< 0.001	0.420	large

Are the model curves consistent with the empirical curves?

It appears that the model curve reflecting high parental difficulties resembles the empirical curve of the second cluster. In contrast, the model curve representing low parental difficulties does not match the empirical curve of the first cluster, as their shapes differ significantly. To assess the degree of fit unambiguously, it is necessary to apply the method for determining curve proximity developed by Aranowska.

The normalized means obtained via cluster analysis for the first cluster were as follows (see Figure 16.3):

$\tilde{u}_{\text{diff}} = 0{,}2$; $\tilde{u}_{\text{rig}} = 0{,}48$; $\tilde{u}_{\text{agr}} = 0{,}16$; $\tilde{u}_{\text{con}} = 0{,}32$; $\tilde{u}_{\text{indif}} = 0{,}12$;

For the second cluster, the following values were obtained:

$\acute{u}_{\text{diff}} = 0{,}45$; $\acute{u}_{\text{rig}} = 0{,}59$; $\acute{u}_{\text{agr}} = 0{,}57$; $\acute{u}_{\text{con}} = 0{,}62$; $\acute{u}_{\text{indif}} = 0{,}50$;

To estimate the fit between curves, two proximity indicators were calculated: η-Aranowska and ζ-Aranowska. For the first cluster, the value of the η-Aranowska indicator was $\eta = -1.02$, while the ζ-Aranowska indicators were respectively $\zeta_1 = 1.089$ and $\zeta_2 = -5.285625$. These results indicate that the first cluster is consistent with the first model curve, which reflects low parental difficulties.

For the second cluster, the value of the η-Aranowska indicator was $\eta = 0.43$, and the ζ-Aranowska indicators were respectively $\zeta_3 = -2.494219$ and $\zeta_4 = 0.193$. The analysis of these results indicates that the second cluster is closer to the second model curve, which reflects high parental difficulties, but does not fully converge with this curve.

The values of the η-Aranowska and ζ-Aranowska indicators were calculated in R using the following code:

```
> x<-c(0.1, 0.3, 0.3, 0.3, 0.3)
> y<-c(0.5, 0.7, 0.7, 0.7, 0.7)
```

```
> c<-c(0.2, 0.48, 0.16, 0.32, 0.12)
> d<-c(0.45, 0.59, 0.57, 0.62, 0.5)
> k=5
> ((sum((c-x)^2))-(sum((c-y)^2)))/(sum((x-y)^2))
[1] -1.02
> ((sum((d-x)^2))-(sum((d-y)^2)))/(sum((x-y)^2))
[1] 0.43
> (var(x)+var(c)-(1/(2*(k-1)))*(sum((c-x)^2)))/(2*var(x))
[1] 1.089375
> (var(y)+var(c)-(1/(2*(k-1)))*(sum((c-y)^2)))/(2*var(y))
[1] -5.285625
> (var(x)+var(d)-(1/(2*(k-1)))*(sum((d-x)^2)))/(2*var(x))
[1] -2.494219
> (var(y)+var(d)-(1/(2*(k-1)))*(sum((d-y)^2)))/(2*var(y))
[1] 0.1932812
```

Table 16.3 Procedure for calculating the indicators for the example of parental difficulties and cold parental mistakes

x	y	c	d	$\Sigma(c-x)^2$	$\Sigma(c-y)^2$	$\Sigma(d-x)^2$	$\Sigma(d-y)^2$	$\Sigma(x-y)^2$
0.1	0.5	0.20	0.45	0.01	0.09	0.1225	0.0025	0.16
0.3	0.7	0.48	0.59	0.0324	0.0484	0.0841	0.0121	0.16
0.3	0.7	0.16	0.57	0.0196	0.2916	0.0729	0.0169	0.16
0.3	0.7	0.32	0.62	0.0004	0.1444	0.1024	0.0064	0.16
0.3	0.7	0.12	0.50	0.0324	0.3364	0.04	0.04	0.16

$$S^2_{e1} = 0.008 \qquad S^2_{e2} = 0.008 \qquad S^2_{r1} = 0.02128 \qquad S^2_{r2} = 0.00483$$
$$\Sigma = 0.0948 \qquad \Sigma = 0.9108 \qquad \Sigma = 0.4219 \qquad \Sigma = 0.0779 \qquad \Sigma = 0.8$$

$$\eta_1 = (\Sigma(c-x)^2 - \Sigma(c-y)^2) / \Sigma(x-y)^2 = (0.0948 - 0.9108) / 0.8 = (-0.816) / 0.8 = \mathbf{-1.02}$$
$$\eta_2 = (\Sigma(d-x)^2 - \Sigma(d-y)^2) / \Sigma(x-y)^2 = (0.4219 - 0.0779) / 0.8 = 0.344 / 0.8 = \mathbf{0.43}$$

$$\zeta_1 = (S^2_{e1} + S^2_{r1} - \tfrac{1}{2}(k-1)\cdot\Sigma(c-x)^2) / 2\cdot S^2_{e1}$$
$$= (0.008 + 0.02128 - 0.125\cdot 0.0948) / 0.016$$
$$= (0.02928 - 0.01185) / 0.016 = \mathbf{1.089375}$$

$$\zeta_2 = (S^2_{e2} + S^2_{r1} - \tfrac{1}{2}(k-1)\cdot\Sigma(c-y)^2) / 2\cdot S^2_{e2}$$
$$= (0.008 + 0.02128 - 0.125\cdot 0.9108) / 0.016$$
$$= (0.02928 - 0.11385) / 0.016 = \mathbf{-5.285625}$$

$$\zeta_3 = (S^2_{e1} + S^2_{r2} - \tfrac{1}{2}(k-1)\cdot\Sigma(d-x)^2) / 2\cdot S^2_{e1}$$
$$= (0.008 + 0.00483 - 0.125\cdot 0.4219) / 0.016$$
$$= (0.01283 - 0.0527375) / 0.016 = \mathbf{-2.494219}$$

$$\zeta_4 = (S^2_{e2} + S^2_{r2} - \tfrac{1}{2}(k-1)\cdot\Sigma(d-y)^2) / 2\cdot S^2_{e2}$$
$$= (0.008 + 0.00483 - 0.125\cdot 0.0779) / 0.016$$
$$= (0.01283 - 0.0097375) / 0.016 = \mathbf{0.1932812}$$

The values of the fit indicators show that the empirical curve of the first cluster is clearly close to the first model curve, which represents low parental difficulties and a low level of parental mistakes. In turn, the empirical curve of the second cluster shows only moderate similarity to the second model curve, which represents high parental difficulties. These results contradict the intuitive assumption regarding the shape of the curves, which would suggest the opposite conclusions.

If the fit between the curves were assessed solely "by eye", the conclusions might be completely different, potentially leading to serious interpretative errors. An intuitive assessment of the curve shapes would suggest that the empirical curve of the first cluster does not align with the first model curve, while the empirical curve of the second cluster would be much closer to the second model curve. However, the fit indicators clearly demonstrated that the reality is different. This highlights the importance of using objective measures of fit, such as the η-Aranowska and ζ-Aranowska indicators, which allow for precise assessment of the correspondence between empirical and model curves, eliminating the risk of subjective errors.

Thus, the computational method constitutes an indispensable tool in the process of verifying curve compatibility, especially in situations where intuition may lead to incorrect or oversimplified interpretations.

Second Example

The second example discusses curve fitting, in which one of the model curves is also an empirical curve. The second study was conducted on a group of 80 parents of early school-age children (7–11 years). The sample included 70 mothers and 10 fathers. The parents' mean age was 36 years; the same value was obtained for the median, and the mode was 35 years. The study involved three groups of respondents: a) the parents completed questionnaires regarding parental mistakes they had committed, b) their children also completed questionnaires concerning the parental mistakes committed by their parents, c) the teachers of these children assessed the level of socio-emotional competence development in the children.

The aim of the study was to determine whether there are relationships between the level of socio-emotional competence development and the level of parental mistakes. Theories concerning parental mistakes equate such mistakes with the acquisition by children of experiences that are unfavourable to their development. Thus, it should be assumed that a higher level of parental mistakes is associated with a lower level of socio-emotional competence development in children.

Figure 16.4 presents model curves illustrating the relationship between parental mistakes and the level of socio-emotional competence development. The model curves are simultaneously empirical curves obtained from the parent sample.

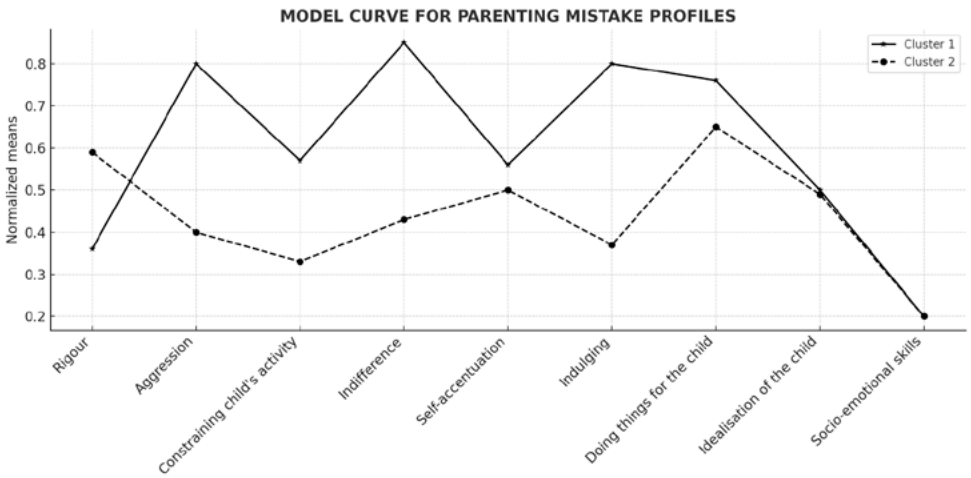


Figure 16.4. Profile curve for the relationship between parental mistakes and the level of socio-emotional competence development – parent group

The standardised means calculated using cluster analysis for the first cluster were as follows (cf. Figure 16.4):

$\tilde{u}_{rig} = 0{,}36$; $\tilde{u}_{agr} = 0{,}80$; $\tilde{u}_{con} = 0{,}56$; $\tilde{u}_{indif} = 0{,}85$; $\tilde{u}_{self\text{-}a} = 0{,}56$; $\tilde{u}_{indul} = 0{,}80$; $\tilde{u}_{doing} = 0{,}77$; $\tilde{u}_{ideal} = 0{,}51$; $\tilde{u}_{soc\text{-}em} = 0{,}20$

For the second cluster, the following values were obtained:

$\mathring{u}_{rig} = 0{,}59$; $\mathring{u}_{agr} = 0{,}40$; $\mathring{u}_{con} = 0{,}33$; $\mathring{u}_{indif} = 0{,}43$; $\mathring{u}_{self\text{-}a} = 0{,}51$; $\mathring{u}_{indul} = 0{,}37$; $\mathring{u}_{doing} = 0{,}65$; $\mathring{u}_{ideal} = 0{,}50$; $\mathring{u}_{soc\text{-}em} = 0{,}20$

The results obtained from the parent sample indicate that the level of parental mistakes—whether high or low—does not significantly differentiate the level of socio--emotional competence development in children. Cluster analysis revealed that the second cluster consists of parents who commit parental mistakes less frequently, and this group accounts for 20% of the study sample. In contrast, the first cluster includes parents who commit parental mistakes more frequently, encompassing as much as 80% of the study participants (Table 16.4).

Table 16.4 Cluster means, number of cases within clusters, and percentage of cases within clusters in the analysis of children's socio-emotional skills and parental mistakes – parent group

	Cluster 1	Cluster 2
Rigour	14.109	17.250
Aggression	18.875	14.438
Constraining child's activity	15.312	12.312
Indifference	25.859	19.500
Self-accentuation	18.797	17.750
Indulging	12.281	8.312
Doing things for the child	17.578	15.750
Idealisation of the child	15.703	15.312
Socio-emotional skills	7.406	7.375
Number of cases	64	16
Percentage (%)	80	20

Statistically significant differences between the two clusters were found for all the analysed variables (Table 16.5). The effect sizes of the differences between clusters were substantial for such characteristics as rigour, aggression, constraining the child's activity, indifference, and indulging, where the effects can be classified as large. For the remaining variables, the effect sizes were small or moderate.

Table 16.5 ANOVA results for clusters in the analysis of children's socio-emotional skills and parental mistakes – parent group

Variable	Between-group variance	*df*	Within-group variance	*df*	*F*	*p*-value	η^2	η^2 Interpretation
Rigour	126.253	1	651.234	78	15.121	< 0.005	0.162	large
Aggression	252.050	1	386.937	78	50.808	< 0.005	0.394	large
Constraining child's activity	115.200	1	635.187	78	14.146	< 0.005	0.153	large
Indifference	517.653	1	621.734	78	64.942	< 0.005	0.454	large
Self-accentuation	14.028	1	1483.359	78	0.737	0.393	0.009	very small
Indulging	201.612	1	192.375	78	81.745	< 0.005	0.511	large
Doing things for the child	42.778	1	418.609	78	7.970	0.006	0.092	moderate
Idealisation of the child	1.953	1	1798.796	78	0.084	0.772	0.001	very small
Socio-emotional skills	0.012	1	975.187	78	0.000	0.975	0.000	very small

The empirical curves that served as the basis for fitting the model curves are presented in Figure 16.5.

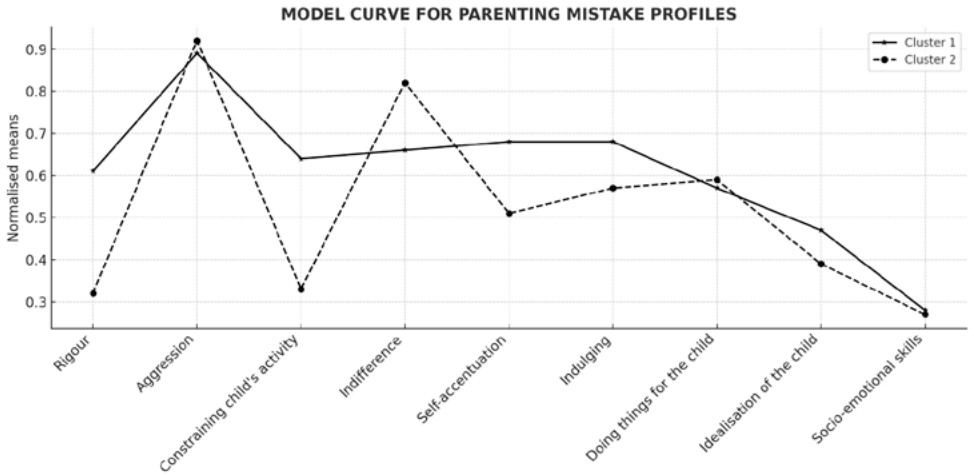


Figure 16.5. Profile curve for the relationship between parental mistakes and the level of socio-emotional competence development – child group

The standardised means calculated using cluster analysis for the first cluster are as follows (cf. Figure 16.5):

$\hat{c}_{rig} = 0{,}61$; $\hat{c}_{agr} = 0{,}89$; $\hat{c}_{con} = 0{,}65$; $\hat{c}_{indif} = 0{,}69$; $\hat{c}_{self\text{-}a} = 0{,}69$; $\hat{c}_{indul} = 0{,}59$; $\hat{c}_{doing} = 0{,}59$; $\hat{c}_{ideal} = 0{,}42$; $\hat{c}_{soc\text{-}em} = 0{,}21$

For the second cluster, the following values were obtained:

$\check{c}_{rig} = 0{,}31$; $\check{c}_{agr} = 0{,}92$; $\check{c}_{con} = 0{,}33$; $\check{c}_{indif} = 0{,}82$; $\check{c}_{self\text{-}a} = 0{,}52$; $\check{c}_{indul} = 0{,}61$; $\check{c}_{doing} = 0{,}45$; $\check{c}_{ideal} = 0{,}36$; $\check{c}_{soc\text{-}em} = 0{,}17$

The results obtained from the child sample indicate that higher levels of parental mistakes committed by parents are associated with lower levels of socio-emotional competence development in their children. Cluster 1 includes 48.755% of the sample, representing the group of parents who, according to their children's assessment, commit more parental mistakes. Cluster 2 comprises 51.25% of the sample and represents the group of parents who, in the opinion of their children, commit fewer parental mistakes. Children from this group demonstrate a higher level of socio-emotional competence development (Table 16.6).

Table 16.6 Cluster means, number of cases within clusters, and percentage of cases within clusters in the analysis of children's socio-emotional skills and parental mistakes – child group

	Cluster 1	Cluster 2
Rigour	18.564	14.390
Aggression	6.436	6.634
Constraining child's activity	9.872	6.049
Indifference	16.308	18.439
Self-accentuation	16.590	14.268
Indulging	14.692	15.122
Doing things for the child	15.590	13.805
Idealisation of the child	10.333	9.049
Socio-emotional skills	7.821	7.000
Number of cases	39.0	41.0
Percentage (%)	48.75	51.25

Statistically significant differences between the two clusters were found for all analysed variables, as presented in Table 16.7. The effect sizes of the differences were substantial for such variables as rigour and constraining the child's activity, for which the effects can be classified as large. For the remaining variables, effect sizes ranged from very small through small to moderate, indicating their lesser importance in differentiating between the analysed clusters.

Table 16.7 ANOVA results for clusters in the analysis of children's socio-emotional skills and parental mistakes – child group

Variable	Between-group variance	*df*	Within-group variance	*df*	*F*	*p*-value	η^2	η^2 Interpretation
Rigour	348.204	1	659.345	78	41.192	< 0.005	0.346	large
Aggression	0.785	1	83.101	78	0.737	0.393	0.009	very small
Constraining child's activity	292.126	1	384.261	78	59.297	< 0.005	0.431	large
Indifference	90.794	1	890.405	78	7.953	0.006	0.092	moderate
Self-accentuation	107.715	1	723.484	78	11.612	< 0.005	0.130	moderate
Indulging	3.689	1	840.697	78	0.342	0.560	0.004	very small
Doing things for the child	63.675	1	551.874	78	8.999	< 0.005	0.103	moderate
Idealisation of the child	32.980	1	882.569	78	2.914	0.092	0.036	small
Socio-emotional skills	13.456	1	961.743	78	1.091	0.299	0.014	small

Using the method developed by Aranowska, it was verified whether the curve plotted in the parent sample corresponded to the curve plotted in the child sample. The results of the fit indices revealed that the curve describing Cluster 1 in the child group is more similar to the curve describing Cluster 1 in the parent group ($\eta_1 = -0.459$, $\zeta_1 = 0.730$, $\zeta_2 = 0.621$). Similarly, the curve describing Cluster 2 in the child group is also more similar to the curve describing Cluster 1 in the parent group ($\eta_2 = -0.598$, $\zeta_3 = 0.962$, $\zeta_4 = 1.087$).

The interpretation of these results indicates that, in both groups, Cluster 1 corresponds to a higher level of parental mistakes, which suggests agreement between parents and children in the assessment of these mistakes. When parents report a greater number of parental mistakes, children also perceive them as more pronounced. Conversely, Cluster 2—associated with a lower level of parental mistakes—also demonstrates agreement: children rate their parents' mistakes as lower when parents themselves evaluate them as less significant.

These results suggest that children are capable of accurately assessing the parental mistakes of their parents, and their assessments are largely consistent with the parents' self-evaluations. This demonstrates that, despite potential differences in perspective, there is significant convergence in the perception of parental mistakes between parents and children, which reinforces the conclusion regarding the validity of parental self-assessment and the accuracy of children's perception.

```
> x<-c(0.36,0.80, 0.56, 0.85, 0.56, 0.80, 0.77, 0.51, 0.20)
> y<-c(0.59, 0.40, 0.33, 0.43, 0.51, 0.37, 0.65, 0.50, 0.20)
> c<-c(0.61, 0.89, 0.65, 0.69, 0.69, 0.59, 0.59, 0.42, 0.21)
> d<-c(0.31, 0.92, 0.33, 0.82, 0.52, 0.61, 0.45, 0.36, 0.17)
> k=9
> ((sum((c-x)^2))-(sum((c-y)^2)))/(sum((x-y)^2))
[1] -0.4587797
> ((sum((d-x)^2))-(sum((d-y)^2)))/(sum((x-y)^2))
[1] -0.597578
> (var(x)+var(c)-(1/(2*(k-1)))*(sum((c-x)^2)))/(2*var(x))
[1] 0.7301463
> (var(y)+var(c)-(1/(2*(k-1)))*(sum((c-y)^2)))/(2*var(y))
[1] 0.620577
> (var(x)+var(d)-(1/(2*(k-1)))*(sum((d-x)^2)))/(2*var(x))
[1] 0.9624022
> (var(y)+var(d)-(1/(2*(k-1)))*(sum((d-y)^2)))/(2*var(y))
[1] 1.087157
```

The detailed procedure for calculating the indicators is presented in Table 16.8.

Table 16.8 Procedure for calculating indicators in the example of parental mistakes and children's socio-emotional skills

x	y	c	d	$\Sigma(c-x)^2$	$\Sigma(c-y)^2$	$\Sigma(d-x)^2$	$\Sigma(d-y)^2$	$\Sigma(x-y)^2$
0.36	0.59	0.61	0.31	0.0625	0.0004	0.0025	0.0784	0.0529
0.80	0.40	0.89	0.92	0.0081	0.2401	0.0144	0.2704	0.1600
0.56	0.33	0.65	0.33	0.0081	0.1024	0.0529	0.0000	0.0529
0.85	0.43	0.69	0.82	0.0256	0.0676	0.0009	0.1521	0.1764
0.56	0.51	0.69	0.52	0.0169	0.0324	0.0016	0.0001	0.0025
0.80	0.37	0.59	0.61	0.0441	0.0484	0.0361	0.0576	0.1849
0.77	0.65	0.59	0.45	0.0324	0.0036	0.1024	0.0400	0.0144
0.51	0.50	0.42	0.36	0.0081	0.0064	0.0225	0.0196	0.0001
0.20	0.20	0.21	0.17	0.0001	0.0001	0.0009	0.0009	0.0000

$$S_{2e1} = 0.050 \qquad S_{2e2} = 0.019 \qquad S_{2r1} = 0.0359 \qquad S_{2r2} = 0.0609$$
$$\Sigma = 0.2059 \qquad \Sigma = 0.5014 \qquad \Sigma = 0.2342 \qquad \Sigma = 0.6191 \qquad \Sigma = 0.6441$$

$$\eta_1 = (\Sigma(c-x)^2 - \Sigma(c-y)^2) / \Sigma(x-y)^2 = (0.2059 - 0.5014) / 0.6441$$
$$= (-0.2955) / 0.6441 = \mathbf{-0.458}$$
$$\eta_2 = (\Sigma(d-x)^2 - \Sigma(d-y)^2) / \Sigma(x-y)^2 = (0.2342 - 0.6191) / 0.6441$$
$$= -0.3848 / 0.6441 = \mathbf{-0.598}$$

$$\zeta_1 = (S_{2e1} + S_{2r1} - \tfrac{1}{2}(k-1)\cdot\Sigma(c-x)^2) / 2\cdot S_{2e1}$$
$$= (0.050 + 0.0359 - 0.0625\cdot 0.2059) / 2\cdot 0.050$$
$$= (0.0859 - 0.01287) / 0.1 = \mathbf{0.730}$$

$$\zeta_2 = (S_{2e2} + S_{2r1} - \tfrac{1}{2}(k-1)\cdot\Sigma(c-y)^2) / 2\cdot S_{2e2}$$
$$= (0.019 + 0.0359 - 0.0625\cdot 0.5014) / 2\cdot 0.019$$
$$= (0.0549 - 0.0313375) / 0.038 = \mathbf{0.621}$$

$$\zeta_3 = (S_{2e1} + S_{2r2} - \tfrac{1}{2}(k-1)\cdot\Sigma(d-x)^2) / 2\cdot S_{2e1}$$
$$= (0.050 + 0.0609 - 0.0625\cdot 0.2342) / 2\cdot 0.050$$
$$= (0.1109 - 0.0146375) / 0.1 = \mathbf{0.962}$$

$$\zeta_4 = (S_{2e2} + S_{2r2} - \tfrac{1}{2}(k-1)\cdot\Sigma(d-y)^2) / 2\cdot S_{2e2}$$
$$= (0.019 + 0.0609 - 0.0625\cdot 0.6191) / 2\cdot 0.019$$
$$= (0.0799 - 0.03869) / 0.038 = \mathbf{1.084}$$

Summary and Discussion

The proposed method of curve comparison enables precise determination of whether a model curve is fitted to empirical curves and to what extent. The results obtained using the η-Aranowska and ζ-Aranowska coefficients allow for the assessment of

curve similarity, as well as for the identification of cases in which the curves differ significantly. By applying this curve-fitting method, it was shown that the perception of parental mistakes by children with a lower level of socio-emotional skills was closer to Cluster 1 of the parent group—those who committed more parental mistakes. At the same time, no cluster was found in the child group that resembled Cluster 2 of the parent group. This result indicates that children rated their parents' mistakes higher than the parents rated themselves. Such conclusions could only be drawn through the use of the curve-fitting method.

The cluster analysis results emphasise the importance of comparing curves across all variables rather than only selected ones. For example, the fact that the aggression mistake in Cluster 1 of the parent group is at a similar level as in Cluster 1 of the child group does not in itself mean that the two clusters (curves) are similar. Without a comprehensive comparison of the curves, it is easy to fall into the illusion of similarity based on one or two variables, while ignoring differences in the others.

The coefficients used offer many advantages but also some limitations. One key issue is the expansion of the range of fit coefficient values as the number of variables in the profiles increases, which may significantly hinder interpretation. In practice, this means that the longer the profile (e.g. 12 or 20 traits), the lower the capacity of classical methods to determine whether a given individual truly follows the model structure or merely fits a random pattern.

This limitation was overcome thanks to a new profiling method proposed by Szymańska, which transfers the analysis into a Hilbert space with a reproducing kernel (RKHS) using RBF kernel functions (Szymańska, 2025b). The key innovation of this method lies in transforming classical feature profiles into geometric points in a relational space, where each point reflects not the values of variables but the similarity of a given individual to others in the dataset (Szymańska, 2025b).

In this approach, the psychological profile gains a new dimension: its length (i.e. number of traits) not only does not limit the analysis but in fact allows for a more accurate reflection of the relational cognitive structure — the person vector can be of any length, and each new transformation in the kernel function takes into account the full configuration of traits. As a result, model fit analysis becomes possible not at the level of traits but at the level of the topology of the person space.

Szymańska developed a new transformation of the classical η-Aranowska coefficient into a kernel-based version, denoted as $\boldsymbol{\eta}^{\star}$(*eta prime*), which enables determination of whether a given individual lies closer to the model profile or to the empirical profile, defined as the geometric mean of the so-called support vectors that determine the cognitive structure of the sample in RKHS space (Szymańska, 2025b).

Thanks to this method, profiling becomes numerically unlimited—both in terms of the number of variables and the number of individuals—because the analysis is conducted not in the observation space, but in a geometric space where each point contains information about an individual's relations to the entire system. This marks a radical shift: from the space of variables to the space of individuals, from feature analysis to the analysis of geometric configurations. Such an approach opens up new

possibilities not only for psychology, but also for artificial intelligence systems operating at the level of relational matching—where it is not the number of features, but their mutual similarities that define functional cognitive profiles. In the future, this may play a key role in the development of AI agents capable of learning not only through variable-based analysis but also through relational mapping between cognitive profiles in kernel spaces.

It is therefore worth asking whether, within the scope of classical two-dimensional models, there are other approaches that could compete with Aranowska's method without the need to transition into kernel space (i.e. Hilbert space). An alternative may be offered by **Latent Growth Curve Models (LGM)**, which are sometimes used to plot empirical curves. However, LGM belong to the class of structural models and require meeting certain assumptions, such as the normality of variable distributions (Dornyei & Otto, 1998; Heck & Thomas, 2009). In the case of psychological measurement scales with skewed results, LGM may not be suitable.

Cluster analysis based on data mining algorithms has no such limitations. It can be applied to variables with diverse characteristics, regardless of scale length, result distribution, or inter-variable relationships. As a method belonging to the data mining class, it allows the inclusion of many variables in the analysis, in accordance with Big Data principles (Elder et al., 2012; Nisbet et al., 2009; Szymańska, 2017c). As such, cluster analysis offers greater flexibility in psychological research and other fields of the social sciences.

16.2. Supplementing the Solution of Structural Equation Models with Cluster Analysis: Applications in Scientific Research

The approach presented in this chapter constitutes an original solution that integrates structural equation models (SEM) with cluster analysis. This type of integration was developed in response to the need for a deeper understanding of the relationships represented in SEM models, which are not always intuitively clear to all users.

Integrating SEM with cluster analysis brings significant benefits, as it allows for the visualisation of the strength of relationships between variables and the identification of groups of individuals in the studied sample—and, consequently, also in the population. This approach not only facilitates the interpretation of results but also enables a better understanding of the data structure in the context of the analysed variables. For this reason, it is particularly valuable in studies that require both precise statistical analyses and a clear presentation of findings.

Structural equation models (SEM) are widely used for modeling complex relationships between variables due to their ability to account for multiple dependencies simultaneously. This is one of their greatest strengths, as it allows for a detailed analysis of the strength of relationships between components of the studied structure

(Szymańska, 2016b). However, although SEM provides information on the direction and strength of relationships between variables, it does not allow for the identification of individual group membership based on the intensity of those variables. This is where cluster analysis becomes highly useful.

Cluster analysis enables the grouping of individuals based on the similarity of their results, allowing for the identification of specific subgroups within a population. As a result, it offers deeper insight into the data structure and its interpretation. The combination of SEM and cluster analysis allows for a more comprehensive picture of the phenomena under study. SEM provides information about the relationships between variables, while cluster analysis makes it possible to identify and characterise groups based on those variables, significantly enhancing analytical capabilities.

This methodological integration yields numerous benefits. SEM provides detailed information on the structure of dependencies within a population but does not indicate how these dependencies are distributed across different groups. Cluster analysis fills this gap, allowing for the visualisation of groups that differ in the intensity of variable characteristics. Consequently, it becomes possible to precisely determine which groups of individuals are more vulnerable to specific issues and how these issues manifest in various population segments.

The combination of SEM and cluster analysis also has practical applications, enabling the creation of more effective interventions and support programmes. As a result, the analysis outcomes become more accessible, data visualisations are easier to interpret, and the communication of research findings—both in academic settings and in practice—is more effective. Thanks to the synergy of these methods, studies gain in precision, comprehensiveness, and practical utility.

SEM is a statistical technique that enables the modeling of complex relationships between variables by using covariance matrices to estimate model parameters. This method allows for the simultaneous inclusion of multiple dependencies within a single analysis, which makes it possible to precisely understand the structure of the phenomena under investigation. The results of SEM are indicators of the strength and direction of the relationships between variables, which are crucial for analysing complex systems.

Cluster analysis, on the other hand, is a statistical technique used to group observations in such a way that units within a single cluster are more similar to each other than to those in other clusters. One of the most commonly used methods in cluster analysis is the ***k*-means method**, which involves dividing the data into k clusters. The aim of this method is to minimise within-cluster variance while maximising between-cluster variance.

The combination of SEM and cluster analysis involves the complementary application of these two techniques. SEM is first used to determine the strength and direction of relationships between variables, followed by the use of cluster analysis to understand how these relationships are distributed across different groups within the studied population. This kind of integration enables better identification of population segments that are most at risk of particular problems and allows for the analysis of how these problems manifest within different groups.

To better understand how the integration of SEM and cluster analysis works, reference may be made to a study conducted on a sample of mothers of preschool-aged children, described in Appendix C. These mothers responded to questions regarding their applied parental goals, experienced parenting difficulties, the mental representation of the child in the parent's mind, and ways of responding to difficult parenting situations, such as withdrawal or applying pressure to the child. Additional questions addressed the constraining of the child's activity and communication style, including aggressive directiveness.

The results of the structural equation model in this study revealed significant relationships among the analysed variables. The relationship between the discrepancy in parental goals and experienced parenting difficulties was high, amounting to 0.75. A similarly strong relationship (0.84) was found between parenting difficulties and the representation of the child in the parent's mind. A negative representation of the child was, in turn, strongly associated with the use of aggressive directiveness (0.7). Moderate correlations were observed between the use of pressure and constraining the child's activity, amounting to 0.6 and 0.64 respectively (Figure 16.6). These results indicate the existence of strong and moderate relationships between variables, where an increase in one variable is associated with an increase in the other.

Despite its precision, SEM does not provide information on how these relationships are distributed across different groups within the studied population. To address this and to understand the distribution of these relationships, cluster analysis was applied, allowing for the identification of specific groups within the sample. As a result, a more detailed picture of the analysed phenomena can be obtained.

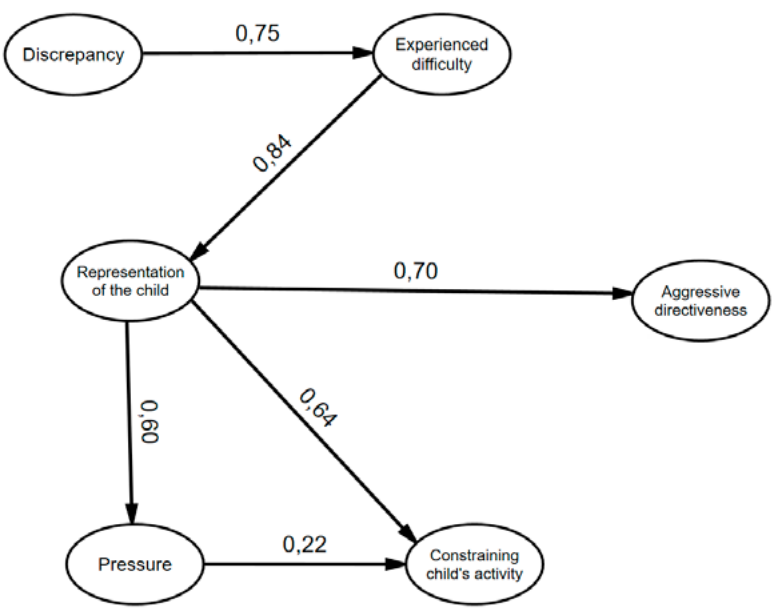


Figure 16.6. Fragmentary structural model of parental mistakes

Figure 16.7 presents the results of the cluster analysis conducted using the *k*-means method for the example under discussion. The algorithms identified two clusters: the first cluster, marked in blue, includes 108 individuals, which constitutes 68.78% of the sample; the second cluster, marked in red, includes 49 individuals, or 31.21% of the sample (Table 16.9).

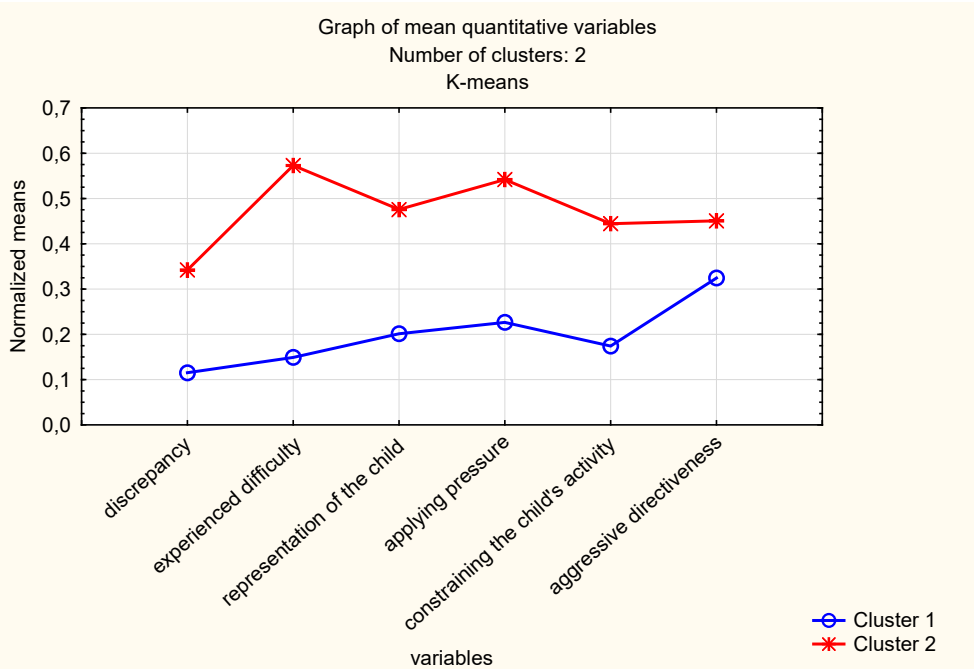


Figure 16.7. Plot of mean quantitative variables (discrepancy, experienced difficulty, representation of the child, use of pressure, constraining child's activity, aggressive directiveness) – Cluster analysis using the *k*-means method

Table 16.9 Mean values of variables for two clusters (discrepancy, experienced difficulty, representation of the child, use of pressure, constraining child's activity, aggressive directiveness) – *k*-means method

Cluster	1	2
Discrepancy	123.407	366.020
Experienced difficulty	11.917	45.837
Representation of the child	11.269	26.653
Use of pressure	6.565	15.714
Constraining child's activity	32.380	82.673
Aggressive directiveness	41.435	49.388
Number of cases	108.0	49.0
Percentage (%)	68.79	31.21

The differences between the clusters proved to be statistically significant for all analysed variables (Table 16.10). A key piece of information provided by cluster analysis is the number of elements assigned to each cluster. The results indicate that approximately two-thirds of the sample belong to Cluster 1, while about one-third belong to Cluster 2. Cluster 1 includes individuals who scored low to moderate across all analysed variables, whereas Cluster 2 comprises individuals with moderate to elevated scores.

Cluster analysis also provided valuable insights into the scope of the phenomena examined using structural equation models. The results clearly show that approximately 30% of the sample experience high levels of parenting difficulties. These difficulties lead to the use of pressure, aggressive forms of communication, and constraining the child's activity.

Importantly, the analysis also revealed a positive conclusion—the majority of the population studied do not experience such difficulties in their relationships with their children. Parents in Cluster 1, for the most part, do not apply pressure, do not constrain their child's activity, and do not exhibit aggressive forms of communication. This is an extremely important and optimistic finding, which sheds new light on parenting relationships in the studied group, providing valuable data that were previously unknown.

Table 16.10 ANOVA results for quantitative variables in cluster analysis (discrepancy, experienced difficulty, representation of the child, use of pressure, constraining child's activity, aggressive directiveness)

	Between-group SS	***df***	**Within-group SS**	***df***	***F***	***p*-value**
Discrepancy	1984030	1	3559653	155	86.3918	<0.005
Experienced difficulty	38782	1	20391	155	294.8003	<0.005
Representation of the child	7978	1	12944	155	95.5305	<0.005
Use of pressure	2822	1	5001	155	87.4634	<0.005
Constraining child's activity	85261	1	96378	155	137.1206	<0.005
Aggressive directiveness	2132	1	6556	155	50.3983	<0.005

Cluster analysis brings an added value that structural equation models (SEM) alone do not provide. While SEM reveals the relationships between variables and explains the strength of these associations, thereby confirming theoretical assumptions in real-world data, it does not answer a key question: what proportion of the studied population is affected by these theoretical assumptions, and which segments of the population are most vulnerable to the phenomena described by the theory? This question is effectively addressed by cluster analysis.

Cluster analysis enables precise identification of which groups of individuals are characterised by specific variable patterns. This makes it possible to understand which population segments are most susceptible to particular problems and how these problems manifest in their everyday functioning. This is particularly

important in the context of designing interventions and support programmes. Such interventions can be more effective when they are precisely tailored to the specific needs of the identified groups within the population, allowing for more efficient problem-solving.

In summary, the combination of structural equation models and cluster analysis constitutes an exceptionally powerful research tool. SEM provides detailed information on the strength and direction of relationships between variables, while cluster analysis enables the identification and characterisation of groups based on these variables. This methodological integration provides a more comprehensive and in-depth understanding of the phenomena under study. The result is the ability to formulate more precise conclusions, implement more effective interventions, and apply research findings more efficiently in practice.

PART IV

Predictive Algorithms and Machine Learning: Theory and Applications

Predictive algorithms and machine learning methods constitute key tools in modern data analysis, finding applications across various disciplines, including psychology, pedagogy, and the social sciences. Their essence lies in the ability to model complex relationships that are difficult to capture using traditional statistical methods. Through the use of algorithms such as artificial neural networks (ANN), support vector machines (SVM), the naïve Bayes classifier, or the *k*-nearest neighbours algorithm (k-NN), it becomes possible to create highly accurate analytical models tailored to the specificity of psychological data.

Part Four of this book focuses on the theoretical foundations of these methods and their applications within the context of the social sciences. Both the operational principles of the algorithms and their implementation in empirical research will be discussed. The structure presented enables the reader to gain insight into the general assumptions of these techniques as well as the specific methods for applying them in research practice.

Artificial neural networks, inspired by the functioning of the brain, make it possible to uncover complex patterns in data, and are therefore used in behavioural analysis, personality modeling, or assessing the effectiveness of therapeutic interventions. Support vector machines are advanced tools for classification and regression that use transformations into higher-dimensional spaces, enabling a more accurate representation of nonlinear relationships. The naïve Bayes classifier and the *k*-nearest neighbours algorithm, although simpler in their assumptions, offer effective solutions for analysing problems that require pattern recognition in data characterised by a high degree of variability.

Within this part of the book, the reader will find not only a detailed description of the operation of each algorithm but also their application in specific research cases. Each chapter includes a theoretical analysis, a review of the literature, and a methodological discussion, allowing for a comprehensive understanding of the capabilities these techniques offer. Particular attention will be devoted to their potential for verifying theoretical models, analysing both quantitative and qualitative data, and forecasting outcomes in psychological contexts.

In summary, this part of the book is intended for researchers and practitioners who wish to understand the foundations of predictive algorithms and learn how to apply them effectively. This material provides tools for critical analysis, research design, and inference, with an emphasis on methodology and scientific precision.

CHAPTER 17

Foundations of Artificial Neural Networks

Neural networks have long played a significant role across various scientific disciplines. They are applied in fields such as materials science, medicine, and particularly in the development of expert systems (Sokołowski & Kosmol, 1995; Cholewa, 1995; Luger & Stubbledfield, 1989; Michalik, 2006a; Wołowiec-Korecka, 2016). A substantial body of literature describes these technologies (Duch et al., 2000; Osowski, 1994; Rutkowski, 2006; Tadeusiewicz, 1993, 2001; Tadeusiewicz et al., 2007; Żurada et al., 1992).

Neural networks are mathematical models inspired by the structure and functioning of biological brains. They consist of artificial neurons that process information and learn from data. Traditionally, neural networks are used for pattern recognition, image analysis, natural language processing, and many other tasks.

In recent years, neural networks have been increasingly applied in psychology. Examples of such use include studies showing that neural networks can predict personality traits and sexual orientation based on appearance (Segalin et al., 2017; Y. Wang & Kosinski, 2018). The analysis of complex patterns in psychological data using these technologies opens new and intriguing research possibilities.

Neural networks can be used to construct predictive models capable of forecasting human behaviour, therapeutic outcomes, and even clinical diagnoses. For instance, these models may analyse data from psychological interviews, psychometric tests, and behavioural observations in order to generate accurate predictions.

The aim of this chapter is to present the potential contributions that neural networks can make to psychology, with particular emphasis on building predictive models. We will discuss the fundamental principles of neural network operation, their architecture, and various types, such as multilayer perceptrons and recurrent networks. We will also present examples of their applications in psychology and outline future directions of research in this field.

17.1. Operational Aspects of Artificial Neural Networks

Artificial neural networks (ANNs) are applied in many scientific disciplines, such as neurobiology, materials science (Basheer & Hajmeer, 2000; Wołowiec-Korecka, 2016), and others. They enable the modeling of complex relationships, which makes them extremely useful in data analysis and processing. Tadeusiewicz notes that "neural networks can be applied with a high likelihood of success wherever problems arise that require the construction of mathematical models capable of reflecting complex relationships between certain input signals and selected output signals" (Duch et al., 2000).

In order for a network to solve a given problem effectively, the problem must be defined as a function with specified input and output values. In statistical and experimental studies, the input values correspond to explanatory variables (predictors, independent variables, etc.), while the output values represent explained variables (dependent variables).

For example, in a medical study on pneumonia, input values may include data from X-ray images, whereas the output values would indicate the presence of pneumonia (value 1) or absence of the disease (value 0). The task of the neural network is to learn to recognize patterns in the input data that allow for predicting the probability of pneumonia occurrence.

This chapter presents how ANNs learn and includes examples of their application to real psychological data. It will be shown how artificial neural networks can be used in psychology to predict outcomes of variables of interest. The potential of ANNs in the development of psychology will also be discussed, including their role in model validation and the construction of expert systems (Aranowska & Szymańska, 2017).

An artificial neural network (ANN) is a processor composed of simple computational elements. It resembles the human brain in two respects: a) it acquires knowledge from the environment through the process of learning, and b) it stores the acquired knowledge. This knowledge is stored in the connection weights between neurons. These weights reflect the strength and nature of the connections, which enables the network to efficiently process and retain information.

A fundamental feature of ANNs is their ability to generalize acquired knowledge to new, previously unseen patterns (also referred to as the network's ability to approximate the values of a multivariable function). One of the important characteristics of neural networks is their ability to process information in parallel, which distinguishes them from traditional algorithms that process data sequentially (Tadeusiewicz, 1993).

The prototype of all neural networks is the human brain. A neural network is a simplified model of the human brain. There are two types of ANN architectures: feedforward networks, in which signals flow in one direction only, and feedback networks, which include recurrent connections (e.g., the Hopfield network).

Neurons in networks can be connected according to three principles: full interconnection (each-to-each), inter-layer connections (in layered networks), or selective connections, e.g., only between neighbouring neurons. In order for the network to generate patterns correctly, it must contain an appropriate number of neurons. A multilayer network consists of at least three layers: a) an input layer, which holds information about the input variables; b) a hidden layer, which contains artificial neurons; and c) an output layer, which holds information about the output variables. The network may have one or more outputs and can be used to solve classification problems (for discrete data) or regression problems (for continuous data).

During the learning process, the network may encounter the so-called *overfitting* problem, which occurs when the network learns specific cases too well and fails to generalise. This problem arises when there are too many neurons in the hidden layer (Tadeusiewicz, 2001). In such cases, the network memorises patterns and does not generalise its solutions, leading to a decline in its predictive power, typically manifested by poor performance on the validation set. The problem of overfitting is also characteristic of the human mind. A child may become so "overfitted" to the addition of certain numbers that they do not apply addition rules to other digits, relying solely on memorised examples. When faced with a new problem involving unfamiliar digits, the child may fail to use the rule and give a random answer, making an error. A similar phenomenon occurs in networks. When overtrained, the prediction error of the network increases.

To verify whether the network has correctly recognised patterns and generated rules, its performance is never evaluated on a single dataset. The dataset is randomly divided into several parts: one subset is used for training the network, while the others serve to test the solutions.

17.1.1. Functioning of Artificial Neural Networks

An artificial neural cell (perceptron) operates analogously to a living nerve cell (neuron). Just like a biological neuron, which receives information via dendrites, processes it, and transmits it through an axon to the next nerve cell (see Figure 17.1), the perceptron operates in three steps: it receives information, sums it in the summation block, and then, in the activation block, makes a decision according to inference rules, transmitting the result further (see Figure 17.2).

Figure 17.2 illustrates that input values denoted as x_i (x_1, x_2,…, x_i) are assigned weights (ω_1, ω_2, ω_i), which are summed in the summation block (Σ). These weights function as carriers of information drawn from the database. In the activation block (F), functions enabling the system to learn are applied. This process unfolds as follows: the system receives input values (x_i) from the dataset and, using inference rules, generates a set of principles. These principles construct an inference system, which calculates the output values $y(t)$, leading to the generation of a solution (Rutkowski, 2006).

With knowledge of the target output value $d(t)$, the system modifies the weights (ω_i) in such a way that the resulting output $y(t)$ approximates the target value $d(t)$

(Rutkowski, 2006). Let us now consider how the system learns: how are the rules created? How does the system know which decision to make? What is the meaning of the target output value *d(t)*? What does it mean that the system modifies the weights so that the output *y(t)* approaches the target *d(t)*?

Figure 17.1: Structure of a Human Neuron

(illustration generated by the SORA model, OpenAI)

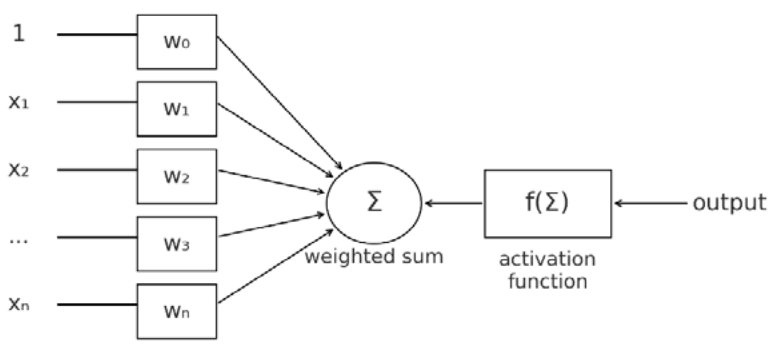


Figure 17.2: Structure of an artificial neuron with multiple input values, where: x – input data, ω – input signal weights, y_t – output function.

Source: author's own elaboration, generated with the assistance of a language model (ChatGPT, OpenAI), based on the diagram available at: https://pl.wikipedia.org/wiki/Neuron_McCullocha-Pittsa

Figure 17.2 presents the perceptron—the first practical model of an artificial neuron, designed by Frank Rosenblatt in 1958. The perceptron expands upon earlier theoretical concepts by McCulloch and Pitts and was the first model capable of learning from data. Its structure—input data x, weights ω, activation function, and output y_t—became the foundation of modern neural networks.

Artificial neural network algorithms learn from information contained in a database. For a learning algorithm, the database functions similarly to the environment for an intelligent being. Intelligent beings acquire knowledge of reality through interactions with their environment. Likewise, artificial intelligence algorithms use input data (x_i) and output data ($d(t)$) from the database, which reflects a portion of reality. The database serves as the learning material from which the algorithm estimates the target output value $d(t)$—for example: how many individuals experienced improved well-being after psychotherapy?

Let us imagine a database containing information on the course of psychotherapy for 50 patients, of whom 40 improved and 10 did not. Information about the target output state $d(t)$ is used by the algorithm during learning. Just as a flight simulator serves as a training tool for a pilot, the database is the learning material for the algorithm. The neural network analyses which input values are associated with the target output value. In the context of psychotherapy, algorithms determine which conditions contribute to positive or negative treatment outcomes.

The algorithm generates rules and models of the phenomenon, for instance: if the therapist worked on transference with the patient, therapy outcomes were positive (rule 1); if the therapist focused solely on current problems, the outcomes were weaker (rule 2) (Grzesiuk et al., 2017). The algorithm assigns such values to the weights (ω_i) that allow the model to best fit the data. Neural networks treat a sample of reality (the database) as a pattern. Just as a pilot practices using a flight simulator, the algorithm repeatedly performs calculations to optimise the weights.

Over time, through trial and error, a human learns how to land in the simulator. This serves as a metaphor for approximating the obtained value $y(t)$ (the trainee pilot lands in the simulator) to the target value $d(t)$ (a correct landing). The algorithm modifies the weights until the value it computes (e.g., predicted classification of patients into successful or unsuccessful therapy groups) approximates the target value (actual group membership of the patients).

For example, the algorithm reproduces the psychotherapy process for patients who benefited and those who did not. It then checks whether it has correctly reconstructed the process by "guessing" whether, based on the indicated differences, it can accurately classify individuals into the group that experienced therapeutic improvement or the one that did not.

Let us now examine how the algorithm operates from a formal perspective. The algorithm retrieves input data (x_i) from the database, multiplies them by their corresponding weights (ω_i), and then sums the results in the summation block according to formula (17.1) (Rutkowski, 2006):

(17.1) $$S = \sum_{i=0}^{n} x_i \, \omega_i$$

where:
x_i – input signal,
ω_i – weight.

As a result of the summation, a signal (S) is generated, which is then processed in the activation block by the activation function (f). The operation of the neuron is expressed by formulas (17.2) and (17.3):

(17.2) $$y(t) = f(S)$$

(17.3) $$y(t) = f\left(\sum_{i=0}^{n} x_i \, \omega_i + \theta\right)$$

where:
$y(t)$ – the approximate output value,
x_i – input signal,
ω_i – weight,
θ – bias (threshold), representing external disturbance.

Knowledge and information are stored in the weights, which can be adjusted through learning. During the learning process, the weights are modified. Initially, input signals ($x_1, x_2, \ldots, x_i$) are provided, for which the corresponding output values ($t = 1, 2, \ldots, j$) are known. These output values are referred to as target output signals ($d(t)$). The set of input data together with the corresponding target values forms the so-called training sequence.

Learning consists of modifying the weights in such a way as to minimise the difference between the target signal ($d(t)$) and the output signal ($y(t)$). The goal is to reduce the error (Θ). The algorithm performs computations until the error is minimised and a satisfactory result is achieved (Rutkowski, 2006). The algorithm's operation process is illustrated in Figure 17.3.

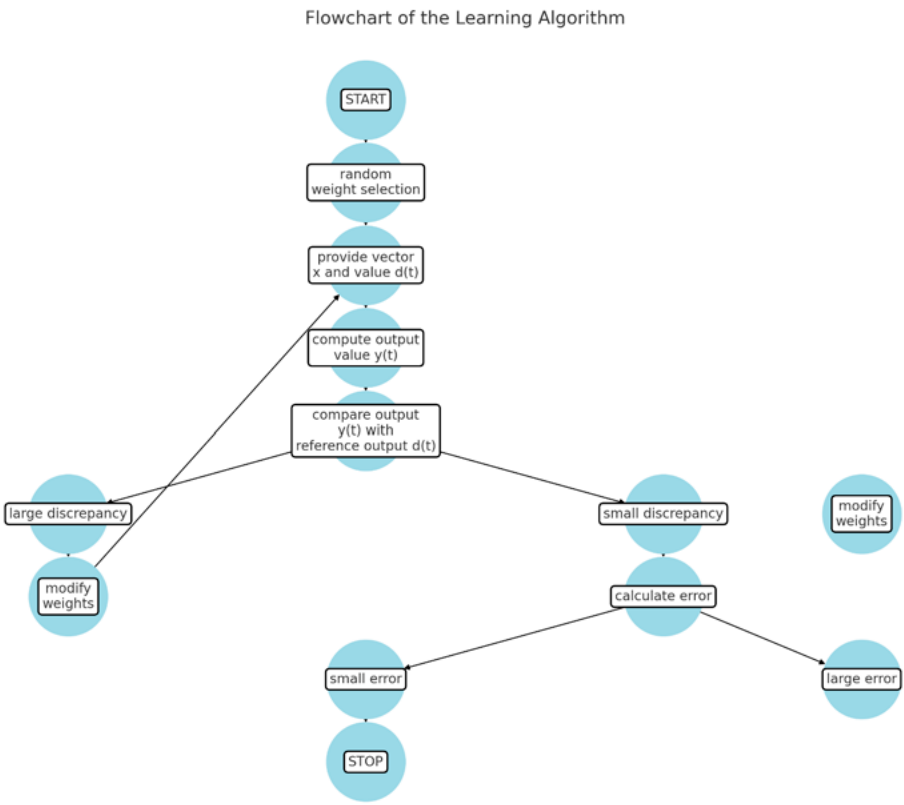


Figure 17.3. Operation process of the algorithm (perceptron)

Source: author's own elaboration, generated with the assistance of a language model (ChatGPT, OpenAI), based on the diagram by Rutkowski (2006).

The algorithm adjusts the weights (ω_i) until they are configured in such a way that the obtained solution is as close as possible to the target value in the dataset (output signal $d(t)$). Much like assembling a puzzle according to a reference image, the algorithm aims to achieve the best possible match.

Some algorithms verify their predictions on a different dataset. This is akin to a pilot training on a different simulator to confirm their skills. These algorithms do not use the entire dataset at once; instead, they divide it into parts: one for training, one for testing the results, and a third for validating the solutions (Elder et al., 2012; Nisbet et al., 2009).

The data are usually randomly divided into three groups—for example, 70%, 15%, and 15% of the observations—although the researcher may choose a different split. The first part, known as the training set, is used by the algorithm for learning. Based on it, the algorithm sets the weights and constructs the rule base. The second part, the test set, allows the assessment of how well the algorithm has learned from the training data and offers an opportunity to adjust the solutions. The third part, the validation set, assesses the accuracy of the algorithm's predictions. On this set, the system verifies

whether the solutions generated from the training and test sets are actually valid—for example, whether it is possible to accurately determine, based on the course of psychotherapy, which patients benefited from the treatment. The validation set is not made available to the algorithm during the learning process. The result obtained on the validation set is the true test of the algorithm's predictive ability.

Despite good performance on the training and test sets, the network may still yield poor results on the validation set. This situation often occurs in the case of network overtraining (Tadeusiewicz et al., 2007). The solutions of an overfitted network cannot be generalised, which significantly limits its usefulness (Nisbet et al., 2009).

17.2. Applications of Artificial Neural Networks in Psychology

In psychology, the use of artificial neural networks (ANNs) is not yet widespread, although their potential could significantly contribute to the development of the field. Currently, there are relatively few publications addressing how ANNs could support psychological research, despite the vast possibilities they offer. Over the past several years, modeling—primarily conducted through structural equation modeling (SEM) and path analysis—has achieved notable success in psychological sciences. These methods are aimed at hypothesis testing, parameter estimation (including the strength of relationships between variables), and evaluating model fit, i.e. their theoretical adequacy (Bartholomew et al., 2008; Hair et al., 2006; Konarski, 2009; Szymańska, 2016b). Meanwhile, comparatively little attention has been paid to predictive models[7], which focus on forecasting individual outcomes.

Nonetheless, algorithms such as decision trees and text mining are being increasingly employed to analyse relationships between variables within datasets (Bokus et al., 2017; Rzechowska, 2004; Szymańska, 2012; Tarwacka-Odolczyk et al., 2014). In psychology, we primarily concentrate on estimating relationships and dependencies between variables, yet we rarely ask ourselves about the practical capacity to predict outcomes based on these models (Szymańska, 2018, 2019). Of course, the value of constructing models is not being questioned. Models are necessary to verify theories and to integrate detailed knowledge into a coherent framework (Tadeusiewicz et al., 2007). Nevertheless, it is also worth reflecting on the extent to which the models we construct—especially those with relatively low levels of association between variables—can be applied in practice. Are they truly capable of effectively supporting prediction and describing reality?

Later in this chapter, we present an example of how a neural network was used to assess the predictive validity of outcomes based on variables included in an SEM

[7] Recent research suggests that predictive coding plays a crucial role as a memory mechanism at the neurochemical level, integrating prediction with neuronal processes responsible for the storage and retrieval of information. This concept has been extensively developed by Karl Friston and colleagues within the framework of predictive coding theory (Barron et al., 2020).

model explaining the inhibition of child activity. Artificial Neural Networks (ANNs) have wide-ranging applications in psychology, enabling novel approaches to the diagnosis and treatment of mental disorders, the analysis of neuroimaging data, and the modeling of cognitive processes. Thanks to advanced machine learning algorithms, ANNs can identify complex patterns in behavioural and brain-related data—an ability that is essential for precision psychology.

Despite the challenges associated with implementing ANNs, their applications are gaining significance, as illustrated by the following examples. One of the key uses of ANNs is the diagnosis and prediction of mental disorders. Bzdok and Meyer-Lindenberg (2017 emphasize that machine learning methods such as neural networks can analyse complex patterns in brain, behavioural, and genetic data, enabling more accurate diagnosis and treatment of mental disorders. Traditional diagnostic categories may not reflect the true causes of disorders, whereas ANNs allow for the identification of biological subgroups of patients, which leads to more personalised therapy (Bzdok & Meyer-Lindenberg, 2017).

The study by Plis and colleagues (2014) revealed that deep learning methods are capable of learning physiologically meaningful representations and detecting hidden relationships in neuroimaging data (Plis et al., 2014). Such analyses contribute to a better understanding of brain functioning and its relationship to behaviour and emotions, which is invaluable in neuropsychological research.

Predictive models are also used in personalised therapy, enabling the development of tailored treatment plans based on patient data analysis. Cohen and DeRubeis (2018) discuss how machine learning methods can predict which therapeutic interventions are most likely to be effective for a given patient. Personalised therapy increases treatment effectiveness and shortens the time required to achieve improvement, which is particularly important in the case of mental disorders such as depression (Cohen & DeRubeis, 2018).

Text and natural language analysis is another area in which predictive models are applied. Calvo and colleagues (2017) demonstrate how natural language processing (NLP) can be used to analyse texts posted on social media platforms such as Facebook and Twitter, enabling the detection of emotions and the identification of individuals who may require psychological support. Such analyses may also be helpful in personalising health interventions by providing valuable information about users' emotional and mental states (Calvo et al., 2017).

In summary, the application of predictive models, including artificial neural networks, in psychology offers numerous benefits. It enables more accurate diagnoses, a better understanding of cognitive and emotional processes, and the personalisation of therapy. The scientific literature available in open access provides numerous examples and detailed studies illustrating these possibilities and their practical applications. ANNs in psychology not only expand our research capabilities but also open new directions for clinical practice, including more precise diagnostics and personalised treatment.

17.3. Case Study: Practical Application of Artificial Neural Networks

Contemporary psychological research increasingly requires the use of advanced data analysis methods that enable both the verification of theoretical assumptions and the practical prediction of outcomes at the individual level. In this context, the combination of structural equation modeling (SEM) with artificial neural networks (ANNs) offers highly promising analytical possibilities that can significantly extend traditional research approaches (Szymańska, 2018, 2019).

The aim of this chapter is to demonstrate how ANNs can complement the traditional use of SEM by providing more advanced tools for the analysis of psychological data. Based on the SEM model presented in the work of Szymańska and Aranowska (2016), along with the associated empirical data, this chapter shows how ANNs can be applied to predict an outcome variable such as child activity inhibition. While the SEM model provides a solid theoretical and empirical foundation, its predictive capacity can be enhanced through the use of ANNs, which are capable of capturing more complex and nonlinear patterns in data.

It is particularly important to highlight that nonlinearity—often overlooked or simplified in classical psychological models—finds its natural expression in ANNs. Psychology, as a discipline long rooted in linear and additive models, rarely incorporates the complex, dynamic structure of nonlinear relationships. Yet such relationships may play a critical role in describing developmental, emotional, and social phenomena. Introducing ANNs as a complement to SEM not only improves the predictive validity of models, but also deepens the understanding of psychological mechanisms that do not conform to linear assumptions.

The traditional SEM model allows for the specification of relationships between theoretical variables; however, its capacity to predict individual outcomes is limited (Szymańska, 2018). ANNs, due to their ability to recognise and analyse complex patterns in data, can provide additional insights not visible in conventional SEM analyses. In this way, integrating these two analytical approaches opens new avenues for research, allowing for more precise data analysis and more accurate outcome predictions.

This chapter begins with a detailed presentation of the SEM model used in the study, outlining its key variables and theoretical relationships. It then presents the application of ANNs to analyse the same data, focusing on their predictive capabilities. Finally, it discusses the differences and potential benefits resulting from the integration of these two approaches, showing how they can mutually complement each other in both theoretical research and practical applications.

Through this approach, the reader will gain insight into the potential offered by advanced data analysis techniques in psychology. The conducted analysis will demonstrate how the integration of SEM and ANNs can lead to a deeper understanding of the studied phenomena, while simultaneously opening new developmental pathways in this scientific domain.

Research Plan

The aim of the study was to determine whether the theoretical constructs identified by Gurycka could be used to predict the level of parental constraining of a child's activity. These theoretical constructs include: the discrepancy between the parental goals[8] and the child's current level of development in the domain of the trained trait; difficulties experienced by parents in upbringing situations; the representation of the child in the parent's mind, consisting in perceiving the child and their tasks as less important than those of the parent; and two defensive responses to stress, namely the use of pressure and withdrawal.

The study used variables derived from Szymańska's model, which constituted a structural reconstruction of Gurycka's theory concerning the emergence of parental mistakes such as the constraining of the child's activity (Gurycka, 1990, 2008). This model is discussed in detail and presented below (Figure 17.4). The study sample is described in Appendix C.

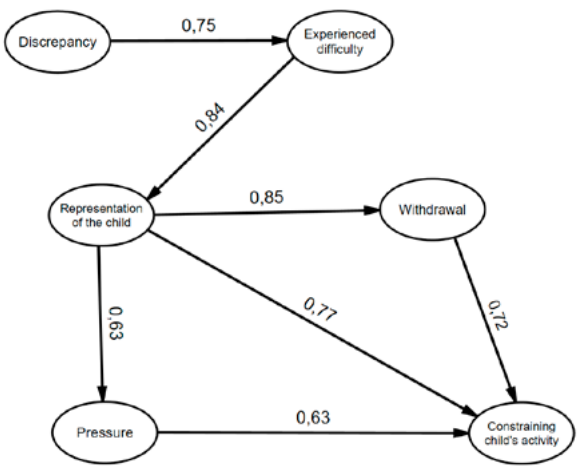


Figure 17.4. Structural model reconstructed on the basis of Gurycka's theory with values of relationships between variables verified by the structural equation modeling procedure

Source: Author's own elaboration.

In her theory (1990), Gurycka assumed that parental mistakes result from the process of coping with a difficult upbringing situation. Gurycka, following Tomaszewski (Tomaszewski, 1975, 1982), defined a difficult situation as one of deprivation, which arises when a parent is unable to reconcile the discrepancy between the traits they would like to develop in the child and those the child is actually developing (Figure

[8] Parenting goals are psychological traits that, during the upbringing process, a parent aims to shape in the child (Brzezińska, 2002; Glenn, 2005; Miller, 1966; Muszyński, 1972; Sośnicki, 1966).

17.4, variable: *Discrepancy*). This discrepancy determines the level of difficulty experienced by the parent (variable: *Experienced difficulty*, Figure 17.4).

As a result of experiencing difficulty, the parent forms a representation of the child that involves perceiving the child's tasks as less important than their own (Figure 17.4, *Representation of the child*). In such a difficult situation, identified by Gurycka—following Reykowski—as a form of stress, the parent may adopt two of the four stress defence reactions indicated by Reykowski: withdrawal or combating stress by means of pressure (Figure 17.4, variables: *Withdrawal* and *Use of pressure*). According to Reykowski (1966), both forms are non-adaptive. Gurycka suggests that the use of pressure, as well as the very emergence of the representation of the child and their tasks as less important than those of the parent, may lead to **constraining the child's activity**. The model also examines whether withdrawal may be associated with **constraining the child's activity**. According to the structural models, the relationships between variables are moderate to strong, ranging from $\boldsymbol{\beta = 0.63}$ to $\boldsymbol{\beta = 0.85}$. Thus, according to Gurycka, a difficult situation—i.e., a situation of deprivation—increases the likelihood of the emergence of a parental mistake in the form of **constraining the child's activity**.

Although the relationships between the variables are moderate to strong, the question remains whether the values of the variable constraining the child's activity can be predicted based on the variables included in the model. If so, this would indicate that they are significant predictors of constraining the child's activity. To answer this question, the method of artificial neural networks was applied.

Input and Output Variables Included in the Construction of the Artificial Neural Network

The input variables for the ANN included: the discrepancy between the traits parents wish to shape in the child and the child's current level of development of those traits; the difficulty experienced by the parents; the representation of the child in the parent's mind; the use of pressure; and withdrawal. The output variable for the ANN was constraining the child's activity.

Results of the Artificial Neural Network Analysis

A total of 200 feedforward neural networks were constructed, from which the best-trained network was selected. The selected network had 5 inputs, 4 neurons in the hidden layer, and 1 neuron in the output layer, and was denoted as MLP 5-4-1. This network structure resulted from the number of input variables, the output variable, and the optimal number of neurons in the hidden layer required for accurate prediction of the output variable (constraining the child's activity). The network could not have too few neurons in the hidden layer, as it would not be able to predict accurately, nor too many, to avoid overfitting (Tadeusiewicz et al., 2007). Figure 17.5 presents the structure of this network.

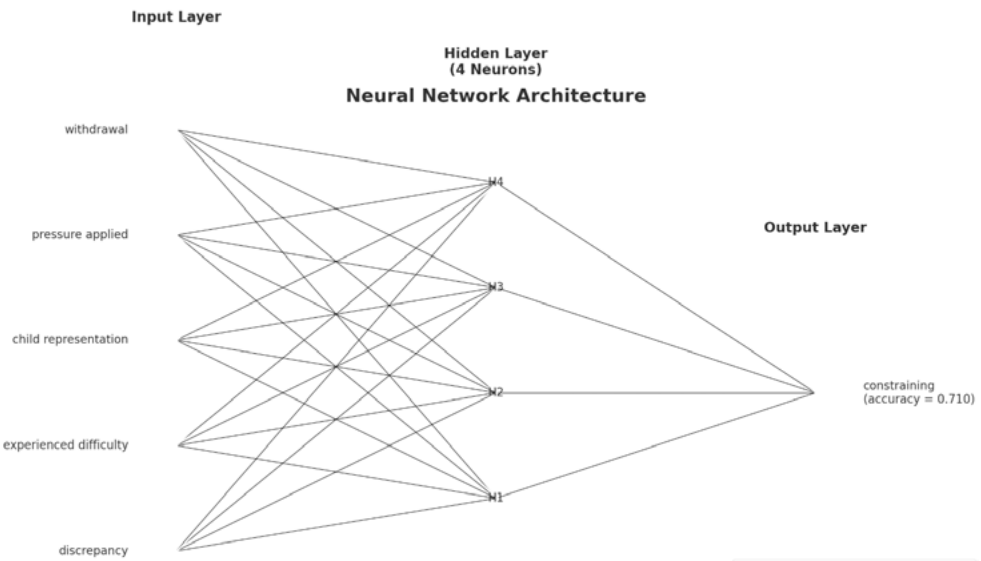


Figure 17.5. MLP 5-4-1 network constructed to determine the constraining of the child's own activity

Source: Author's own elaboration.

The input data for the network included: discrepancy, difficulty experienced in the upbringing situation, the representation of the child in the parent's mind, use of pressure, and withdrawal. During the training process, the patterns were divided into three separate sets: a training set (70% of observations), a testing set (15% of observations), and a validation set (15% of observations). The training set consisted of 223 individuals, accounting for 70% of the sample. The testing set included 48 individuals (15% of the sample), and the validation set also included 48 individuals (15% of the sample). The training set was used to train the network, the testing set was used to verify the network's predictions during the learning process, and the validation set was used to assess the accuracy of the predictions after the learning process had been completed.

The validation set allowed for the evaluation of the accuracy of the predictions for each individual based on the input variables. The accuracy of prediction was determined by the correlation between the network's predicted values and the actual data. The higher the correlation coefficient, the better the network was assessed. Table 17.1 presents a prediction sheet showing sample results for 12 individuals in terms of the input variables: discrepancy, experienced difficulty, representation of the child, use of pressure, and withdrawal, as well as the output variable, which was constraining the child's activity ("input constraint").

Table 17.1. Prediction sheet presenting sample results for 12 individuals in terms of input variables and the output variable predicted by the ANN – constraining.

Discrepancy	Experienced difficulty	Representation of the child	Use of pressure	Withdrawal	Constraining – actual	Constraining – predicted	Residual
430.00	34.00	28.00	20.00	24.00	91.00	84.5641	-6.436
19.00	16.00	2.00	10.00	6.00	24.00	30.7167	6.717
353.00	21.00	5.00	10.00	2.00	45.00	36.9662	-8.034
35.00	12.00	7.00	4.00	12.00	89.00	29.0321	-59.968
247.00	34.00	19.00	10.00	19.00	49.00	55.4609	6.461
7.00	16.00	10.00	24.00	8.00	52.00	54.5165	2.517
289.00	30.00	24.00	11.00	13.00	68.00	54.8097	-13.190
223.00	34.00	4.00	20.00	6.00	47.00	52.1131	5.113
112.00	6.00	26.00	4.00	1.00	20.00	30.4218	10.422
115.00	22.00	15.00	16.00	16.00	49.00	55.3003	6.300
18.00	8.00	5.00	3.00	2.00	38.00	21.9164	-16.084
104.00	19.00	32.00	13.00	13.00	51.00	55.1214	4.121

Source: Author's own elaboration.

In Table 17.1, the column "Constraining – predicted" provides the predictive value for the i-th individual in terms of the predicted level of constraining the child's activity, based on the network's calculations. The last column, "Residual", shows the discrepancy between the expected result (i.e., the network's prediction – "Constraining – predicted") and the actual result obtained by the individual on the test ("Constraining – actual"). The residual was calculated according to the following formula:

$$\varepsilon = y(t) - d(t)$$

Where:
ε – residual
$y(t)$ – value predicted by the network for the t-th individual
$d(t)$ – actual value obtained by the t-th individual (target output value)

Table 17.2 presents a summary of the training process for the network classifying the levels of constraining the child's activity. As shown in Table 17.2, the prediction accuracy of the network was 0.710 for the validation set, 0.723 for the training set, and 0.788 for the testing set.

Table 17.2. Summary of the training process for networks classifying the constraining of the child's activity

Network name	Accuracy (train)	Accuracy (test)	Accuracy (validation)	Error (train)	Error (test)	Error (validation)	Training algorithm	Error function	Activation (hidden)	Activation (output)
MLP	0.723	0.788	0.710	0.009	0.022	0.011	BFGS	SOS	Tanh	Exponential

Source: Author's own elaboration.

17.4. Summary and Conclusions from the Study on Artificial Neural Networks

The calculation results indicate that the input data were effectively processed by the network, allowing for accurate prediction of the values of $y(t)$ corresponding to the level of constraining the child's activity. The relationship between the output values predicted by the network $y(t)$ and the target values $d(t)$ was high and amounted to 0.710. Figure 17.6 presents the values predicted by the network for constraining the child's activity $y(t)$ (y-axis) and the target values $d(t)$ (x-axis) for each individual in a coordinate system. If the network had predicted the results perfectly, all values would align along a straight line.

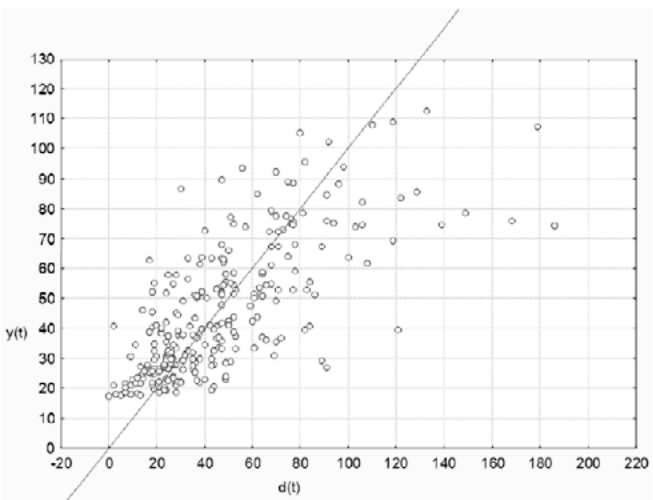


Figure 17.6. Target values $d(t)$ and values $y(t)$ predicted by the network presented in a coordinate system

Source: Author's own elaboration.

The five input values included in the structural model reconstructed on the basis of Gurycka's theory enabled the network to predict individuals' results. This means that the variables: discrepancy, difficulty experienced in the upbringing situation, representation of the child in the parent's mind, use of pressure, and withdrawal from the upbringing situation allow for reliable prediction of parents' results in the domain of constraining the child's activity. This was the main goal of the analyses using the network.

It should be noted that the relationship between the output values predicted by the network $y(t)$ and the target values $d(t)$ in all datasets—training, testing, and validation—was similarly high. For the training set, the correlation was 0.723; for the testing set, 0.788; and for the validation set, 0.710 (Table 17.2). These high and consistent correlation values demonstrate that the network not only predicts the output

variable (constraining the child's activity) well, but more importantly, that the ANN generalizes effectively and transfers learned rules to new datasets that it did not encounter during training.

This provides evidence of the usefulness of Gurycka's theory in explaining the causes of constraining the child's activity. Based on five input values, it is possible to predict individuals' results in this domain with reasonable accuracy.

Let us now examine the result of individual no. 4 (Table 17.1). For this person, the network predicted a value of 29 in terms of constraining the child's activity, while the actual value was 89, meaning the residual was nearly 60 (this individual can be easily found in Figure 17.6). It was found that 26 individuals, or 8.15% of the total sample, had a prediction error of at least 30 points. Of these, the network underestimated the results for 16 individuals and overestimated them for 10.

Such a large underestimation may indicate a certain atypicality in response patterns. As noted by Aranowska (2016), atypical responses affect the reliability of measurement. This poses a challenge for psychological sciences, encouraging revision of scales and closer examination of theoretical constructs (Aranowska, 2016).

The evidence supporting the usefulness of Gurycka's theory in explaining the causes of constraining the child's activity suggests that the input variables are valuable predictors. Nevertheless, the neural network also pointed to cases in which the model does not fully capture the complexity of reality, highlighting the importance of further research into diverse variants of parental experiences.

In this context, strategies such as the Reconstruction of the Transformation Process (RTP) may serve as valuable complements in the analysis of psychological phenomena. As Rzechowska (2004, 2011c) has shown, RTP enables a more detailed description of different developmental variants and allows for the identification of atypical trajectories in the upbringing process (Rzechowska & Szymańska, 2017). Combining data mining methods with psychological theories—such as the RTP strategy—may not only enrich psychological theory but also provide more precise diagnostic and prognostic tools.

The results of this study open new possibilities for psychology, emphasizing the need to integrate traditional theoretical models, such as SEM, with modern analytical technologies, including neural networks and data mining methods.

Conclusions

The application of neural networks in psychological sciences can provide many interesting insights into the usefulness (or lack thereof) of various models reconstructed based on psychological theories. This allows for the identification of models that can be effectively used for prediction and the development of predictive models, as well as the rejection of weaker models which, despite offering certain explanations, contribute little in practice. Aranowska and Szymańska (2017) demonstrated that a neural network can challenge the solution proposed by structural models if that solution is inaccurate (Aranowska & Szymańska, 2017). When the operational validity

of a latent variable is low, the relationships between latent variables in a SEM model may remain high, but the predictions based on such a model will be inaccurate. This issue, discussed by methodologists, concerns the fact that when a latent variable is poorly operationalised, it may be more strongly related to other variables in the model than to its own items (Hair et al., 2006). Therefore, it appears that neural networks may have substantial utility in testing the validity of structural models that reconstruct theoretical assumptions.

17.5. Classification Model of Artificial Neural Networks (ANNs)

This chapter presents methods for testing using artificial classification neural networks. Models tested using classification neural networks are those in which the predictors—that is, the explanatory variables (input signals)—may be either quantitative or qualitative, whereas the explained variable (output signal) is always qualitative. Below, an example of a model calculated with such a network is presented. In the model, the explained variable (output signal) was the educator's opinion, describing how a child behaves in kindergarten ("well-behaved" or "difficult"). The explanatory variables (input signals) in the model were: warm directiveness, aggressive directiveness, the child's demand for obedience, and teaching the child rules of appropriate behaviour. The research sample from which the data for this model were derived is described in Appendix B.

As part of the study, 200 artificial neural networks were constructed, among which the best-trained one was a Radial Basis Function (RBF) network. A summary of the network is presented in Table 17.3. This network had four inputs, thirty neurons in the hidden layer, and two neurons in the output layer. The output neurons corresponded to the levels of the variable "educator's opinion" – "well-behaved" and "difficult". The data were divided into three sets: training, testing, and validation. Each of them constituted 33% of the research sample. The training set was used to set the weights in the neurons, the test set evaluated the effectiveness of the weights during the learning process, and the validation set was used to assess the correctness of the weights after the learning process ended.

RBF networks are characterised by a three-layer structure that includes an input layer, a hidden layer, and an output layer. The input layer receives the input data, which are then processed in the hidden layer. The hidden layer uses radial basis functions, such as the Gaussian function, allowing for the modeling of complex, nonlinear relationships. The output layer uses a linear function that makes the final prediction based on the transformed data from the hidden layer.

The use of Gaussian activation in the hidden layer enables the transformation of input data into a higher-dimensional space, increasing the network's capacity to recognise patterns. Linear activation in the output layer allows for generating predictions that are a linear combination of the values obtained in the hidden layer. This configuration

enables effective modeling of relationships between variables, making RBF networks particularly useful in analyses requiring precise recognition of complex dependencies.

Table 17.3. Summary of the training process for networks classifying children's behaviour in kindergarten

Summary of active networks (play master's thesis in play master's thesis)

Network ID	Network Name	Quality (Training)	Quality (Testing)	Quality (Validation)	Learning Algorithm	Error Function	Activation (Hidden)	Activation (Output)
2	RBF 4-30-2	59.722	70	70	RBFT	SOS	Gaussian	Linear

Based on the results in Table 17.3, the best-fitting neural network for classifying children's behaviour in kindergarten was the RBF 4-30-2 network. This network demonstrated a 70% quality rate in both the testing and validation phases, indicating that the model effectively generalises knowledge to data not included in the training process. The learning algorithm used in this network was RBF, and the activation functions were the Gaussian function in the hidden layer and a linear function in the output layer. The Gaussian function, due to its nonlinearity, enables the model to capture more complex relationships in the input data.

The variable "educator's opinion", which served as the output variable, had two qualitative levels. Based on the analysis of the results, it is possible to observe how accurately the neural network was able to classify children as well-behaved or difficult depending on the values of the input variables.

The first level of the variable "educator's opinion" (well-behaved child) was correctly classified by the artificial neural network in 42.11% of cases, and incorrectly in 57.89%. The low classification accuracy for this group indicates the model's difficulty in distinguishing well-behaved children based on the available input data.

For the second level of the variable "educator's opinion" (difficult child), the network correctly classified 79.41% of cases and incorrectly classified 20.59%. The high classification accuracy for difficult children indicates the model's high sensitivity. This may suggest that input variables such as directiveness or the child's demand for obedience have a stronger impact on identifying difficult children than on identifying well-behaved ones.

Model evaluation in terms of sensitivity and specificity

The model is characterised by high sensitivity (79.41%) in identifying difficult children. Sensitivity in this context refers to the model's ability to correctly recognise positive cases—that is, difficult children. High sensitivity indicates that the model effectively classifies children who require more attention in kindergarten due to behavioural difficulties.

In contrast, the model's specificity is 42.11% and refers to the well-behaved children category. Specificity measures the model's ability to correctly identify negative

cases—that is, children who are not difficult. Low specificity suggests that the model struggles to correctly recognise well-behaved children, leading to an overlabelling of well-behaved children as difficult.

Table 17.4. Summary of the classification of "educator's opinion" during the training of the RBF 4-30-2 model educator_opinion (Classification Summary) (play master's thesis in play master's thesis) Sample: Training

	educator_opinion-1	educator_opinion-2	educator_opinion-All
Total	76	68	144
Correct	32	54	86
Incorrect	44	14	58
Correct (%)	42.105	79.412	59.722
Incorrect (%)	57.895	20.588	40.278

Based on the results in Table 17.4, the best-fitting neural network for classifying children's behaviour in kindergarten (RBF 4-30-2) demonstrated that it can effectively predict difficult child behaviour with an accuracy of 79.41%. The prediction of well-behaved child behaviour was less precise, reaching an accuracy of 42.11%. The overall accuracy of the model amounted to 59.72%.

The model exhibits high sensitivity in identifying difficult children, which means it accurately recognises instances of difficult behaviour based on the available input variables. In contrast, the low specificity in classifying well-behaved children suggests that the model struggles to correctly assign this type of behaviour, potentially resulting in the more frequent mislabelling of well-behaved children as difficult.

Description of Sensitivity Analysis Results

Sensitivity analysis makes it possible to evaluate the impact of individual input variables on the model's prediction outcome. The value of 0.007801 indicates that warm directiveness has the greatest influence on the model's prediction outcome among all input variables. However, this value is relatively low, which suggests that although warm directiveness is the most important variable, its absolute effect on the model's prediction is not substantial. Similarly, the value of 0.004251 for the variable demanding obedience indicates a moderate influence on the model's prediction outcome, though this value is also relatively low. Aggressive directiveness, with a value of 0.003324, has a smaller impact on the prediction outcome compared to warm directiveness and demanding obedience. The smallest influence on the prediction outcome is exerted by the variable teaching rules of appropriate behaviour, with a value of 0.002408.

In summary, the sensitivity analysis for the RBF 4-30-2 neural network showed that warm directiveness is the most important variable influencing the prediction outcome, whereas teaching rules of appropriate behaviour has the least influence. However, the values of all variables are relatively low, suggesting that none of them has a very strong individual effect on the model's prediction outcome. These results suggest that in

order to improve the model's effectiveness, particular attention should be paid to the accurate measurement and inclusion of warm directiveness in the input data.

Table 17.5. Sensitivity Analysis Results for the RBF 4-30-2 Network

Networks – Sensitivity Analysis (play master's thesis in play master's thesis) Sample: Training

Network	warm_directiveness	obedience_demand	aggressive_directiveness	teaching_rules
2. RBF 4-30-2	0.007801	0.004251	0.003324	0.002408

Low sensitivity of the input variables does not mean that the model is unable to make accurate predictions. The RBF 4-30-2 neural network operates on complex interactions between input variables and its ability to model nonlinear relationships. In cases of such low sensitivity values, a key factor influencing predictive effectiveness is the network's capacity to identify shared patterns in the input data, which may be more pronounced in certain classes of the output variable.

For difficult children (high sensitivity), the model likely relies on specific combinations of input variable values that are more distinct and easier for the network to recognise. In the case of well-behaved children (low specificity), the uniformity or lack of characteristic patterns in the input variables may lead to lower prediction accuracy.

It is also worth noting that neural networks such as RBF are capable of capturing dependencies between variables that are not visible in sensitivity analysis. For example, interactions between warm directiveness and aggressive directiveness may play a significant role in prediction, even though the individual effects of these variables are low.

These results suggest that the effectiveness of the model's predictions may be the result of the synergistic effect of all input variables combined with the network's ability to efficiently process information in the hidden layer.

Interpretation of the ROC Curve

The ROC (Receiver Operating Characteristic) curve is a tool used to evaluate the performance of a classification model. It illustrates the relationship between sensitivity and specificity, allowing for the assessment of the model's ability to distinguish between positive and negative classes.

In Figure 17.7, the y-axis represents sensitivity, i.e. the model's ability to correctly identify positive cases (difficult children). Sensitivity indicates the percentage of truly positive cases that were correctly classified by the model. The x-axis represents (1 – specificity), which is the proportion of false positives. Specificity refers to the percentage of negative cases (well-behaved children) that were correctly classified as negative.

The ROC curve shows how well the model classifies positive and negative cases. In an ideal scenario, the ROC curve would begin at point (0,0), then rise vertically to point (0,1), and continue horizontally to point (1,1). Such a shape would indicate that the model has perfect sensitivity and specificity.

In the figure presented, the ROC curve does not reach the ideal line, which suggests that the model has certain limitations in classifying cases. The area under the ROC curve (AUC – Area Under Curve) provides a clear measure of model quality – the larger the area, the better the model's performance. AUC values close to 1 indicate an excellent model, whereas values close to 0.5 reflect performance near random guessing.

Based on the presented ROC curve, it can be concluded that the model has moderate ability to classify both positive cases (difficult children) and negative cases (well-behaved children). The model's sensitivity is higher for the difficult children class, as previously confirmed in the sensitivity and classification analyses. The model's specificity is lower, indicating difficulties in correctly classifying well-behaved children.

The RBF 4-30-2 model demonstrates moderate effectiveness in classifying children as well-behaved or difficult, which is evident both in the sensitivity analysis and on the ROC curve. High sensitivity for the difficult children class and low specificity for the well-behaved children class indicate areas where the model can be further improved.

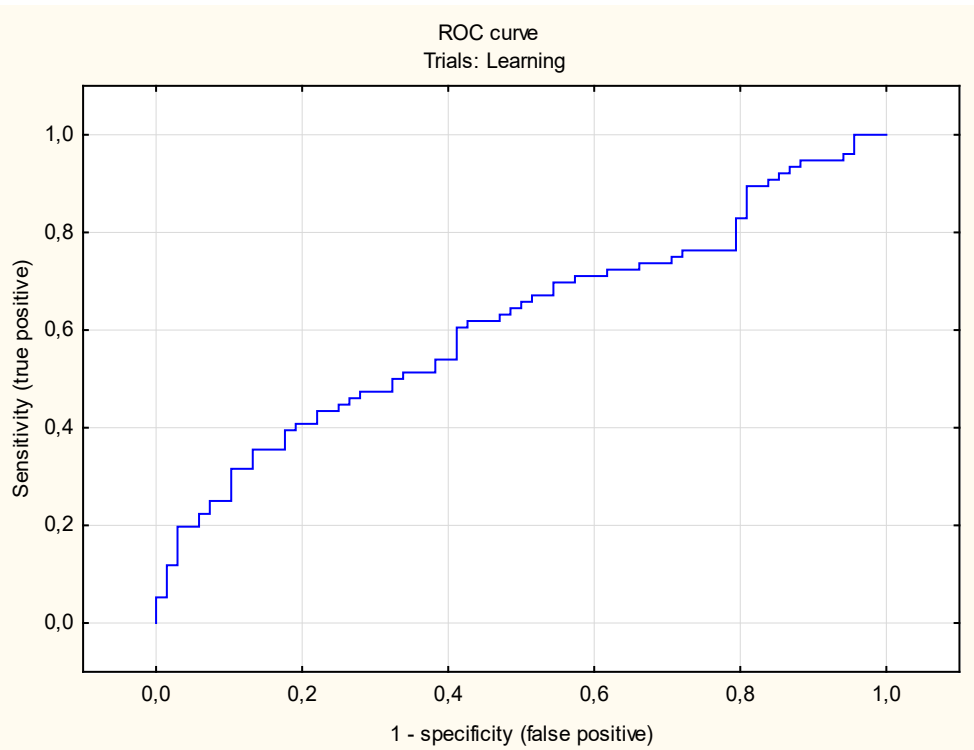


Figure 17.7. ROC curve for the RBF 4-30-2 model

ROC Area and Threshold

The table presents the results of the ROC curve analysis for the RBF 4-30-2 model, showing the area under the ROC curve (AUC – Area Under the Curve) and the ROC threshold. The area under the ROC curve (AUC) is 0.615519. AUC is a measure of the quality of a classification model, and AUC values range from 0 to 1. An AUC value of 0.615519 suggests that the RBF 4-30-2 model performs better than random guessing, though it remains far from perfect. This result indicates a moderate ability of the model to distinguish between classes—that is, well-behaved and difficult children.

The ROC threshold is 0.291336 and represents the cut-off point at which the difference between sensitivity and specificity is maximised. This value indicates the optimal compromise between identifying truly positive cases and minimising false alarms. This means that at this threshold value, the model performs best in distinguishing between positive cases (difficult children) and negative cases (well-behaved children).

Table 17.6 presents the results of the analysis, including the area under the ROC curve and the optimal threshold for the model. These values indicate that although the model demonstrates some classification ability, there is room for further improvement, particularly in terms of specificity.

Table 17.6. ROC Area and Threshold (Samples: Training)

Model	ROC Area (AUC)	ROC Threshold
RBF 4-30-2	0.615519	0.291336

Interpretation of the gain chart

The gain chart (Figure 17.8) illustrates the performance of the classification model in comparison to random selection. The x-axis of the chart represents the cumulative percentage of the population (percentile), while the y-axis shows the cumulative gain, that is, the number of correctly identified positive cases (difficult children). The chart includes two lines: the blue one, representing random selection, and the red one, representing the RBF 4-30-2 model.

The blue line (random selection) serves as a reference point. It shows what the gain would look like under random selection. If the model is no better than random guessing, the gain line for the model will be close to the line of random selection. In contrast, the red line ([2. RBF 4-30-2]) shows the cumulative gain for the RBF 4-30-2 model.

By analysing the chart, it can be observed that the red line runs above the blue line, which indicates that the RBF 4-30-2 model performs better than random guessing in identifying positive cases. The greater the difference between the model line and the random selection line, the better the model's performance. The model

effectively identifies positive cases early on (in the lower percentiles), which means it is more efficient in identifying difficult children in the initial segments of the population. As the percentile increases, the model line approaches the random selection line, suggesting that the model's effectiveness decreases in the higher percentiles of the population.

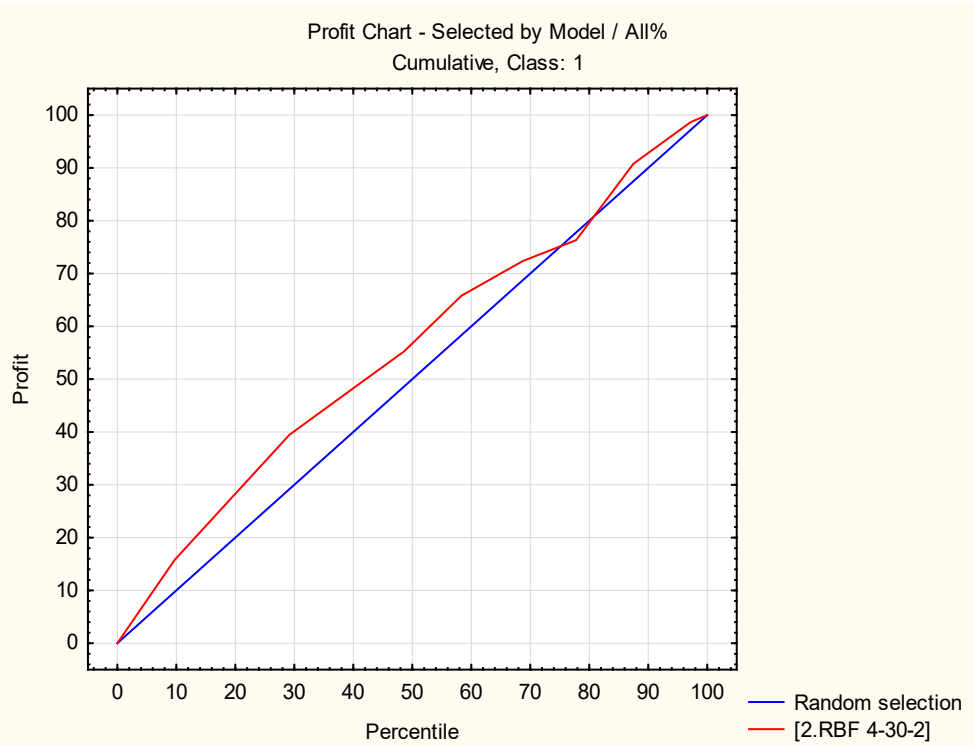


Figure 17.8. Cumulative Gain Chart for the RBF 4-30-2 Model Compared to Random Selection

The conclusions drawn from the analysis of the chart indicate that the RBF 4-30-2 model performs better than random guessing, as evidenced by a higher gain at the beginning of the chart (lower percentiles). However, its effectiveness declines in the higher percentiles, suggesting that the model may be less effective in identifying positive cases across the full population.

In summary, the RBF 4-30-2 model is effective in identifying positive cases (difficult children) better than random selection, particularly in the lower percentiles of the population. Its effectiveness decreases in the higher percentiles, suggesting that the model is more useful for selective identification of positive cases at the beginning of the population.

17.6. Regression Model of Artificial Neural Networks (ANNs)

Regression is one of the key data analysis techniques that enables the prediction of continuous values based on a set of input features. It allows for the prediction of an output value by identifying patterns in the input variables. In the context of artificial neural networks (ANNs), a regression ANN model makes it possible to capture complex, nonlinear relationships in the data. This model is particularly useful when both input and output variables are numerical in nature—that is, when they are interval or ratio variables.

An example of applying a regression ANN model is the analysis of parental upbringing difficulties in relation to the child's temperamental traits. In the discussed model, the dependent variable was the level of upbringing difficulty, and the independent variables were the child's temperamental traits: general activity, activity during sleep, approach, flexibility, positive mood, regularity of sleep, regularity of eating, regularity of habits, concentration, and persistence.

The model consisted of 10 inputs (the child's temperamental traits), 28 neurons in the hidden layer, and one output (level of upbringing difficulty). The data were divided into three sets: training set, test set, and validation set. During the experiment, 200 different networks were trained, among which the best RBF 10-28-1 network achieved the following results: 0.574 accuracy on the training set, 0.400 on the test set, and 0.612 on the validation set.

Table 17.7. Summary of the training process for the regression network predicting experienced upbringing difficulty

Network Name	Training Quality	Testing Quality	Validation Quality	Training Error	Testing Error	Validation Error	Training Algorithm	Error Function	Hidden Activation	Output Activation
RBF 10-28-1	0.574	0.400	0.612	89.122	88.388	79.135	RBFT	SOS	Gaussian	Linear

Explanation of the Term RBF

RBF, or Radial Basis Function network, is a type of artificial neural network used in both classification and regression problems. The main characteristic of the RBF network is the use of radial basis functions as activation functions in the hidden layer. These functions are nonlinear and symmetric in shape, which enables effective modeling of complex, nonlinear relationships in data (Dash et al., 2016).

In an RBF network, each unit in the hidden layer computes the distance between the input and a specific centre, and then applies a radial basis function (most often a Gaussian function) to that distance. The results of these computations are then passed to the output layer, which is linear and combines the results from the hidden layer to produce the final prediction.

RBF networks are particularly effective in tasks where data are nonlinearly separable, as their structure allows for efficient modeling of such complex relationships.

Due to these properties, RBF networks are a valuable tool in data analysis and are used in various scientific and industrial fields where precise modeling of nonlinear relations is required.

Explanation of the Term MLP

MLP, or Multi-Layer Perceptron, is one of the most popular types of artificial neural networks, used across a wide range of tasks, including both classification and regression (Gaudart et al., 2003). The structure of an MLP consists of at least three layers: an input layer, one or more hidden layers, and an output layer.

Each neuron in an MLP is connected to neurons in adjacent layers through weights, which are optimised during the training process. The output value of each neuron is the result of processing input signals through an activation function such as sigmoid, tanh, or ReLU (Rectified Linear Unit). The activation function introduces nonlinearity into the model, enabling it to capture complex relationships in the data.

The training process of an MLP is conducted using the *backpropagation* algorithm, which optimises the network's weights by minimising an error function—e.g. mean squared error (MSE) for regression or cross-entropy for classification. Backpropagation involves computing error grades for each weight and updating them in the direction that minimises the error (Tadeusiewicz et al., 2007).

MLPs are particularly effective in tasks where the data are complex and nonlinearly separable. Their ability to model intricate dependencies makes them widely applicable in fields such as image recognition, text analysis, financial forecasting, and medicine. Although MLPs can be prone to overfitting, the use of regularisation techniques such as dropout and cross-validation significantly improves their generalisation capability on new data.

The multi-layer perceptron is a fundamental tool in the arsenal of modern machine learning methods. It combines structural simplicity with computational power, making it effective in solving many complex analytical problems.

Key Differences Between RBF and MLP Networks

The fundamental difference between RBF (Radial Basis Function) networks and MLP (Multi-Layer Perceptron) networks lies in how they process data and learn the relationships between inputs and outputs. RBF networks use radial basis functions as activation functions in the hidden layer, which allows them to effectively capture local patterns in the data. In contrast, MLP networks employ various activation functions such as sigmoid, tanh, or ReLU, which introduce nonlinearity and enable the modeling of global patterns in the data.

Local patterns in the data are those that hold significance within a limited range of input variables and are more specific to certain regions of the data space. In statistical terms, this can be compared to specific characteristics of a study sample, where the

data reflect unique features of a given group that are not necessarily representative of the broader population (Nisbet et al., 2009). For example, when studying consumer behaviour in a particular geographical region, one may observe that residents of that region have specific purchasing preferences not found in other areas. An RBF network is well suited for modeling such local patterns, as its radial basis functions can effectively capture relationships specific to particular subsets of data.

Global patterns, on the other hand, refer to general relationships present throughout the entire dataset and reflect overarching trends. MLP networks are more effective in modeling such global patterns because their layered structure and variety of activation functions allow them to capture complex, nonlinear relationships across the full range of input data. For example, in public health research, global patterns may involve general relationships between lifestyle and health that apply to the entire population rather than to specific groups.

The reason why an RBF network yields better predictions in one case and an MLP in another depends on the specifics of the problem and the structure of the data. If the data exhibit clear local patterns, an RBF network may produce better results due to its ability to model local dependencies. In contrast, when the data involve more complex, global relationships, an MLP network may be more effective due to its flexibility and capacity to model nonlinear relationships at multiple levels.

In practice, the choice between an RBF and an MLP network depends on the nature of the particular problem, the data structure, and the requirements for prediction. Often, the best approach is to experiment with both architectures and choose the one that yields better results in a given context.

Therefore, in light of the obtained results, the superior prediction performance of the RBF network indicates that the dataset contained significant local patterns (specific to the study sample), which the RBF network was able to effectively capture. This means that the child's temperamental traits exhibited specific dependencies and patterns that were more local and less universal, allowing the RBF network to better fit the model to the data and achieve higher accuracy in predicting the level of upbringing difficulties.

Table 17.8. Sensitivity Analysis of the RBF 10-28-1 Neural Network for Child Temperamental Traits

	Tpositive _mood	Tactiv- ity	Tactivity _sleep	Tflexi- bility	Tpersis- tence	Tap- proach	Tsleep _regular	Thabit _regular	Tatten- tion span	Teating _regular
5. RBF 10-28-1	1.154	1.105	1.048	1.0208	1.016	1.012	1.011	1.009	1.004	1.002

The sensitivity analysis of the RBF 10-28-1 neural network for child temperamental traits shows which traits determine the level of upbringing difficulty experienced by parents. Higher sensitivity values indicate greater importance of a given trait in the model's prediction results.

The highest sensitivity value, 1.154, is associated with the trait "Positive mood". This suggests that a child's positive mood is the most important determinant of the

level of upbringing difficulty. High sensitivity means that changes in this trait significantly affect the model's output, indicating strong importance. The next important trait is "Activity", with a value of 1.105. A child's activity level is also a significant determinant, although not as strong as positive mood, but still playing an important role in predicting upbringing difficulties.

The child's activity during sleep, with a sensitivity value of 1.048, is another important trait. This suggests that changes in sleep activity can significantly influence the assessment of upbringing difficulty. "Flexibility", with a value of 1.0208, also plays a relevant role, although slightly less than the aforementioned traits, but still a meaningful determinant in the model.

Child persistence, with a value of 1.016, and approaching others, with a value of 1.012, have moderate relevance to upbringing difficulties. The child's sleep regularity also plays a role, though moderately, with a value of 1.011. The regularity of habits, with a value of 1.009, is less important compared to the other traits, but still influences the model's predictions.

Child attention span, with a value of 1.004, has little relevance to upbringing difficulties, according to the sensitivity value. The regularity of eating, with a value of 1.002, is the least important trait in the model's output, suggesting that this trait is the least relevant in the context of parenting difficulties.

Strong trait importance in the model's outcome is indicated by higher sensitivity values. The higher the value, the more significant the change in that trait is for the prediction result. In this analysis, values above 1.1 can be considered as indicating strong importance, values around 1.05–1.1 as moderate, and values below 1.05 as low importance.

In summary, temperamental traits such as positive mood, activity, and activity during sleep have the greatest impact on assessing the level of parenting difficulty. In contrast, traits such as eating regularity and concentration have the least impact on the model's prediction result.

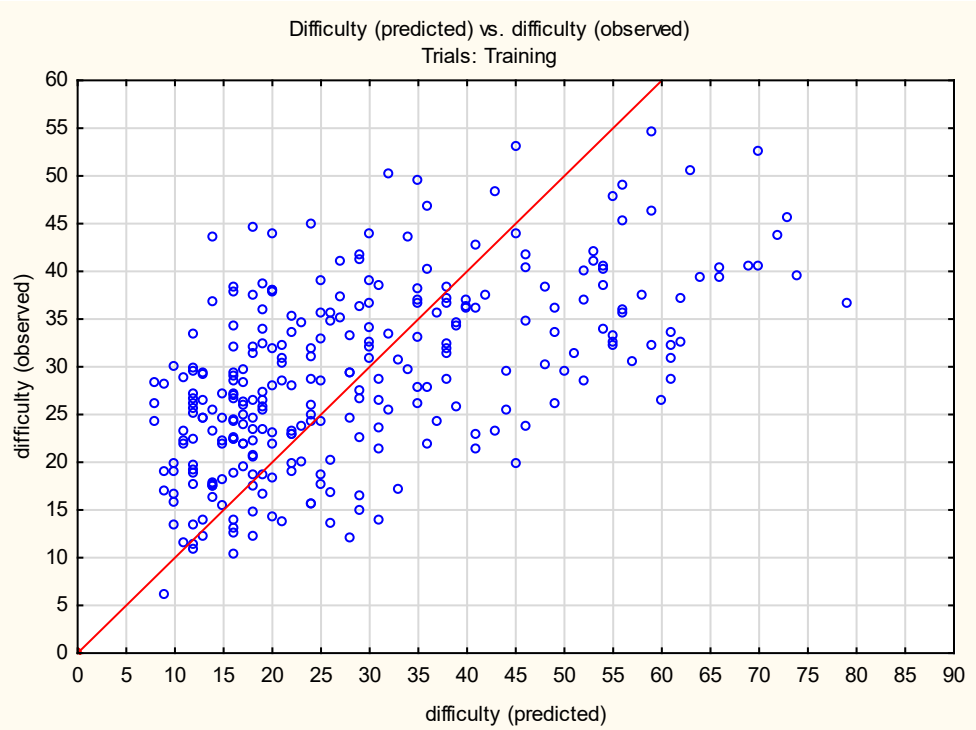


Figure 17.9. Prediction Dispersion of the RBF 10-28-1 Neural Network Compared to Actual Variable Values (difficulty)

The chart illustrates the relationship between the actual values of the dependent variable (difficulty, Actual) and the values predicted by the model (difficulty, Output) for training samples. The X-axis presents the actual values of the dependent variable (Actual), while the Y-axis shows the values predicted by the model (Output).

The points on the chart (blue) represent the values predicted by the network for individual cases. The red line on the chart represents the ideal fit, where the predicted value (Output) equals the actual value (Actual). If the neural network had achieved perfect prediction, all blue points would align along this red line, which is analogous to a linear regression model.

The dispersion of points around the regression line indicates prediction errors. The farther a point is from the red line, the greater the prediction error. This distance is a measure of regression error and is considerable in this case, indicating significant differences between actual and predicted values. In the central part of the chart (for difficulty values from approximately 10 to 40 on the X-axis), most points lie relatively close to the regression line, suggesting that the model predicts the dependent variable more accurately in this range. However, even here some dispersion is noticeable, indicating prediction errors.

Dependent variable values above 40 on the X-axis exhibit even greater dispersion, indicating larger differences between actual and predicted values. At the extreme values of difficulty, the model appears less accurate, and prediction errors become more evident.

For example, for actual results around 70 or 80, the model predicts values around 40 or 45, which means the network significantly underestimates high results. Such dispersion and large prediction errors indicate low model accuracy. The neural network is unable to effectively estimate high values of the dependent variable, which significantly reduces its overall effectiveness.

In summary, the results presented in the chart indicate that the model did not achieve perfect prediction. Large errors, particularly for high values of the dependent variable, show that the neural network struggles to provide accurate forecasts in these areas, which translates into its low predictive accuracy. To make the model more useful, further optimisation is needed to reduce prediction errors, especially for high values of the dependent variable.

These results suggest that the model performs poorly in predicting the level of upbringing difficulty based on the child's temperamental traits. The low accuracy on the test set (0.400) indicates that the model fails to correctly predict the majority of test data. In the context of regression, such an accuracy value may signify considerable differences between predicted and actual values. This situation suggests that the model fails to capture the key relationships between temperamental traits and the level of upbringing difficulty.

Additionally, the significant discrepancy between the results on the training set (0.574), test set (0.400), and validation set (0.612) suggests model instability and difficulties in generalisation. Although this is not a classic case of overfitting—since the highest accuracy occurs on the validation set rather than the training set—such a discrepancy may indicate problems with balancing fit to the data and the ability to predict new cases.

It should also be considered that the child's temperamental traits, despite their widespread recognition in scientific literature, may not be such strong predictors of experienced upbringing difficulties. These results may indicate that other factors, not included in the model, could have a greater impact on the level of upbringing difficulty. This may suggest the need to expand the range of analysed features to include additional variables, such as environmental, socioeconomic factors, or specific family situations.

In summary, these results indicate the necessity for further model optimisation and potentially the inclusion of additional variables in order to improve predictive effectiveness. Without appropriate modifications, this model will not be able to effectively predict the level of upbringing difficulty based on the child's temperamental traits, which may also suggest that temperament is not as strong a predictor as previously assumed.

Other Types of Artificial, Neural Networks (ANNs)

In addition to regression models of artificial neural networks (ANNs), there are many other types of ANNs that are applied in various fields of data analysis. Below are brief descriptions of some of them, with particular emphasis on classification and regression neural networks with time series, as well as clustering analysis performed by neural networks.

Classification neural networks with time series are used for classifying data that are organised in time sequences (Hüsken & Stagge, 2003). Examples of such data include sensor signals, stock market data, or health monitoring system outputs. In these networks, the input data consist of sequences of values that are analysed in the context of their temporal distribution. Popular architectures used for such tasks include recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). These networks are capable of capturing temporal dependencies in the data, which enables effective classification based on patterns occurring over time.

Similar to classification networks, regression neural networks with time series are used to analyse data organised in time sequences. The difference lies in the fact that instead of classifying data into predefined categories, these networks predict continuous values based on temporal patterns. They are widely used in forecasting energy demand, predicting stock prices, analysing climate trends, and many other areas.

Clustering analysis is a data exploration technique that groups datasets based on their similarity. Neural networks can also be used to perform clustering analysis (Du, 2010). One popular type of neural network used in clustering is the Kohonen network, also known as self-organising maps (SOMs). These networks transform input data into lower-dimensional representations, grouping similar data into neighbouring nodes within the network. Clustering analysis performed by neural networks is particularly useful for uncovering data structures, identifying hidden patterns, and segmenting data in a nonlinear way.

In summary, there are many different types of artificial neural networks that are used to solve various analytical problems. Each type of network has its specific applications and advantages that can be utilised depending on the characteristics of the data and the goals of the analysis. Understanding these different types of neural networks enables better alignment of analytical methods with the specific problem, leading to more effective and precise outcomes.

CHAPTER 18

Other Machine Learning Methods

Machine learning is a field of science focused on developing algorithms and models that enable computers to learn and make decisions based on data. In addition to artificial neural networks, discussed in previous chapters, there are many other machine learning methods widely used in data analysis, forecasting, and task automation.

This chapter presents three popular machine learning methods: Support Vector Machine (SVM), Naive Bayes Classifier, and the K-Nearest Neighbours Algorithm (KNN). The theoretical foundations of each method, their applications, as well as their advantages and disadvantages will be described. Each of these techniques has its own unique properties that make it suitable for different applications in data analysis and predictive modeling.

Support Vector Machine (SVM) is a powerful tool primarily used for classification tasks. The fundamental principle of SVM is to find a hyperplane that maximises the margin between different classes in the feature space. This allows for effective separation of classes even in complex and high-dimensional datasets. SVM is particularly valued for its ability to work with data that are not linearly separable, thanks to the use of kernel functions (Evgeniou & Pontil, 2001; Hearst et al., 1998). Examples of SVM applications include image recognition, text analysis, and biomedicine, where precision and efficiency in high-dimensional feature spaces are crucial.

The Naive Bayes Classifier is a simple yet effective algorithm based on Bayes' theorem (Nisbet et al., 2009). It is called "naive" because it assumes that all features are independent, which rarely holds true in practice. Despite this simplification, the Naive Bayes Classifier performs remarkably well in many applications, such as text

classification, spam filtering, and sentiment analysis. Its advantages include speed and efficiency, even with large datasets. However, its drawback is its sensitivity to the independence assumption, which can lead to incorrect results in more complex cases.

The K-Nearest Neighbours (KNN) algorithm is one of the simplest machine learning methods. KNN operates by comparing new data to the k nearest neighbours in the training set and assigning a class based on the similarity of those neighbours (Nisbet et al., 2009). It is a method that does not require a model training phase, which makes it intuitive and easy to understand. The advantage of KNN is its simplicity and effectiveness in both classification and regression tasks, especially when the data naturally form clusters. However, its drawback is high computational complexity with large datasets, as each new case must be compared to all points in the training set. Examples of KNN applications include image recognition, disease diagnosis, and recommendation systems.

In summary, this chapter aims to present key machine learning methods that are widely used in various fields. Each method will be discussed in terms of its theoretical foundations, practical applications, and advantages and disadvantages, enabling the reader to better understand and select appropriate tools for their own analyses.

18.1. Support Vector Machines: Theory and Practice

Support Vector Machine (SVM) is an advanced machine learning method used for both classification and regression. It was developed by Vladimir Vapnik and his collaborators and has become one of the most powerful tools in data analysis (Evgeniou & Pontil, 2001).

In classification, SVM aims to find a hyperplane that best separates samples of different classes in the feature space. A hyperplane is a line in two-dimensional space, a plane in three-dimensional space, or an $(n–1)$-dimensional surface in an n-dimensional space (Evgeniou & Pontil, 2001). The goal of SVM is to maximise the margin, that is, the distance between the nearest data points from both classes and the hyperplane, while minimising classification error. Support vectors are the data points closest to the hyperplane and have the greatest influence on its position, determining the margin and the location of the hyperplane (Hearst et al., 1998).

SVM regression, also known as Support Vector Regression (SVR), is an extension of SVM used for predicting continuous values. SVR operates on a similar principle as SVM for classification, but instead of finding a hyperplane that maximises the margin between classes, SVR looks for a hyperplane that maximises the margin within which most data points are contained. The goal is to minimise prediction error while maintaining model simplicity (i.e. regularisation). The hyperplane is determined so that as many data points as possible fall within a specified margin of error (epsilon). Data points lying outside this margin become support vectors, influencing the shape and position of the hyperplane (Hearst et al., 1998).

Support vectors play a key role in both types of SVM. These are the data points closest to the hyperplane and are crucial to its definition. In classification, they help maximise the margin between classes, whereas in regression they help maintain the hyperplane within the defined error margin (Evgeniou & Pontil, 2001).

Support Vector Machines (SVM) operate by mapping input vectors into a high--dimensional feature space. In this multidimensional space, it is easier to linearly separate data points. This is achieved through the use of a kernel function, which implicitly performs this mapping without the need to directly calculate the coordinates of the data in that space.

This process involves several key stages:

1. **Mapping into feature space:** Input vectors are mapped into a higher-dimensional feature space using a kernel function. This process enables easier linear separation of data that may not be linearly separable in the original input space.
2. **Selection of the optimal hyperplane:** In this feature space, SVM selects a hyperplane that maximises the margin of separation between different data classes. This margin is defined as the distance between the hyperplane and the closest data points from each class.
3. **Determination of support vectors:** The data points located closest to the hyperplane are called support vectors. These points are crucial because they determine the position and orientation of the hyperplane. As stated by Hearst: "The values α_i also have an intuitive explanation. For each training example there is one α_i. Each α_i determines the extent to which that training example influences the SVM function" (Hearst et al., 1998, p. 27).
4. **Solving the optimisation problem:** SVM solves an optimisation problem to find the weight vectors and bias that define the hyperplane. This process involves maximising the margin while simultaneously minimising classification errors. "We can show that the optimal hyperplane is defined as the one with the maximum margin of separation between the two classes" (Hearst et al., 1998, p. 18).
5. **Using the kernel function:** Kernel functions allow computations to be performed in the original input space, avoiding direct calculations of coordinates in the high-dimensional feature space. "We can think of this as a linear algorithm in a high-dimensional space, but in practice it involves no computations in that high-dimensional space. The use of a kernel function allows all necessary computations to be carried out directly in the input space" (Hearst et al., 1998, p. 20).

SVM can use various kernel functions to transform input data into higher dimensions where they can be more easily separated or approximated. The most commonly used kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels. The linear kernel is used when the data are linearly separable, whereas the RBF kernel performs well with nonlinear data (Evgeniou & Pontil, 2001).

The SVM learning algorithm involves solving an optimisation problem that either maximises the margin between classes or minimises the regression error while simultaneously minimising the norm of the weight vector (Evgeniou & Pontil,

2001). The learning process includes preparing the training dataset, selecting an appropriate kernel function, setting model parameters such as the capacity (C) and the error margin parameter (epsilon for SVR), solving the optimisation problem to find the best hyperplane, and evaluating the model on a test set to assess its effectiveness and generalisation ability.

SVM and SVR are widely applied in various fields, such as image classification, bioinformatics, and time series prediction (Evgeniou & Pontil, 2001). In image classification, SVM is used for object recognition, face detection, and handwritten text recognition, whereas in bioinformatics it is employed for DNA sequence classification and protein structure prediction.

Example of an SVM Classification Model

This analysis examined whether the level of experienced upbringing difficulty could be explained by the way a parent responds to stress. The Support Vector Machine (SVM) method was applied for the dependent variable, which took on two values: 1 representing low scores and 2 representing high scores. The independent variables were two continuous predictors: stress responses and the use of pressure and withdrawal.

The aim of the analysis was to determine whether the level of parenting difficulty experienced by parents could be effectively predicted based on their stress responses, manifested through withdrawal and pressure strategies. Support Vector Machines (SVM) are powerful classification tools that enable effective separation of classes in complex datasets. The use of this method allows for precise assessment of whether the independent variables determine the dependent variable.

The analysis aimed to verify whether there is a significant relationship between the stress responses employed by parents and the level of upbringing difficulty they experience. The results of this analysis would provide insight into whether the independent variables are key determinants of the experienced upbringing difficulty. If the SVM model demonstrates high classification accuracy, it indicates that the explanatory variables play a significant role in classifying the level of upbringing difficulty. On the other hand, low model accuracy might suggest that other factors not included in the analysis may be more influential in differentiating between levels of parenting difficulty.

The SVM model configuration included the choice of binary classification, which is appropriate for a dependent variable with two classes. The radial basis function (RBF) kernel was selected, as it is effective for modeling nonlinear relationships in data. The parameter C, which controls the model's capacity, was set to 10, allowing a compromise between maximising the separation margin and minimising classification errors. The number of iterations was set to 1000 to ensure sufficient time for model convergence and identification of the optimal solution.

Additionally, to ensure that the model generalises well to new data, 10-fold cross-validation was conducted. This technique evaluates the model's performance

on different subsets of the data, minimising the risk of overfitting and providing more reliable results. With these settings, the SVM model was tailored to predict the level of upbringing difficulty based on parents' stress responses and use of pressure, enabling a more precise analysis of the relationships between these variables.

The analysed data were drawn from the sample described in Appendix C, which included children aged 3 to 6 years. The sample was divided into a training set comprising 92 cases and a test set containing 28 cases, yielding a total of 120 cases. The results of the SVM model indicate high effectiveness in classifying the experienced level of upbringing difficulty.

Before conducting the analysis, the variable representing experienced upbringing difficulty was standardised, and the analysis included cases that fell below and above half a standard deviation (S) to the left and right of the mean. A value of 1 indicated individuals with scores below half a standard deviation from the mean, while a value of 2 indicated individuals with scores above half a standard deviation. Thus, the total number of individuals included in this analysis was 120, as the remaining individuals—those whose scores fell within half a standard deviation on either side of the mean—were excluded. Appendix C notes that the total sample comprised 319 individuals, but only 120 met the criteria for inclusion in this analysis.

The SVM model, using an RBF kernel and a gamma parameter of 0.5, achieved a classification accuracy of 84.783%. The classification accuracy for the training set was 84.783%, for the test set 85.714%, and for the combined dataset 85.000%. The number of support vectors was 54, of which 52 were bound, indicating the complexity of the model and its ability to separate the classes. The support vectors were evenly distributed between the classes, with 27 for class 1 and 27 for class 2.

These results suggest that stress responses—particularly the use of pressure and withdrawal—are significant determinants of experienced upbringing difficulty. The high model accuracy on both datasets confirms that the SVM model generalises well and effectively classifies levels of upbringing difficulty, which may have practical applications in assessing and supporting parents in challenging parenting situations. A decision rule can be formulated as follows: if a parent tends to employ pressure or withdrawal strategies in stressful situations, they are highly likely to experience an elevated level of upbringing difficulty.

Table 18.1. SVM Model Predictions for the Test Sample

Case ID	Stress Reaction (Predictor)	Pressure & Withdrawal (Predictor)	Upbringing Difficulty (Actual)	Upbringing Difficulty (Predicted)	Accuracy	Confidence Class 1	Confidence Class 2
1	14.0	40.0	2	2	Incorrect	0.0	1.0
6	23.0	31.0	2	2	Correct	1.0	0.0
7	2.0	56.0	1	1	Correct	1.0	0.0
8	31.0	24.0	2	2	Correct	1.0	0.0
18	2.0	40.0	1	1	Incorrect	0.0	1.0
33	0.0	34.0	2	2	Correct	1.0	0.0
37	0.0	43.0	2	2	Correct	1.0	0.0
43	0.0	34.0	1	1	Correct	1.0	0.0
46	12.0	32.0	1	1	Correct	1.0	0.0
62	9.0	23.0	1	1	Correct	1.0	0.0
73	4.0	23.0	1	1	Correct	1.0	0.0
74	3.0	34.0	2	2	Correct	1.0	0.0
76	1.0	38.0	2	2	Correct	1.0	0.0
87	24.0	24.0	2	2	Correct	1.0	0.0
91	1.0	55.0	2	2	Correct	1.0	0.0
96	19.0	20.0	2	2	Correct	1.0	0.0
114	4.0	1.0	2	2	Correct	1.0	0.0
125	0.0	51.0	2	2	Correct	1.0	0.0
138	12.0	0.0	1	1	Correct	1.0	0.0
166	11.0	36.0	1	1	Correct	1.0	0.0
180	20.0	30.0	1	1	Correct	1.0	0.0
188	1.0	38.0	2	2	Correct	1.0	0.0
213	2.0	50.0	1	1	Correct	1.0	0.0
215	0.0	25.0	1	1	Correct	1.0	0.0

Table 18.1 presents the prediction results of the SVM model for the test sample, which included children aged 3 to 6 years. Each row represents an individual case from the test sample, containing the values of the predictors (stress response and the use of pressure and withdrawal), the actual value of the dependent variable (experienced upbringing difficulty), the value predicted by the model, and the classification accuracy.

The column "Experienced Upbringing Difficulty (Predicted)" shows the predicted values provided by the SVM model. The "Accuracy" column indicates whether the prediction was correct (labelled as "Correct") or incorrect (labelled as "Incorrect"). The confidence level for each class is presented in the last two columns: "Confidence Level – Class 1" and "Confidence Level – Class 2", indicating the model's confidence in assigning a case to class 1 or class 2, respectively.

The model achieved high classification accuracy for most cases. The results indicate that the model effectively predicts the level of experienced upbringing difficulty based on parents' stress responses and their use of pressure. In cases where the prediction was incorrect, the model assigned a high confidence level to the wrong class, which suggests the need for further analysis and possible adjustment of model parameters.

The consistently high confidence scores across most cases demonstrate the model's strong certainty in its predictions, which is a positive indicator. However, incorrect predictions with high confidence suggest potential areas for improvement, such as fine-tuning model parameters or considering additional variables that may influence upbringing difficulties.

In summary, the table provides detailed insights into the effectiveness of the SVM model in predicting experienced upbringing difficulty based on parents' stress responses and use of pressure. These results may be useful in practical applications, supporting the assessment and assistance of parents in challenging parenting situations.

Example of a Support Vector Regression (SVR) Model

In the conducted SVR (Support Vector Regression) analysis concerning experienced parenting difficulties, appropriate settings were applied to optimize the predictive model. The data were drawn from the sample described in Appendix A and concerned the same cases that were analyzed using an Artificial Neural Network (ANN) model presented in Section 17.6.

The dependent (explained) variable was the level of parenting difficulties experienced by parents, while the independent (explanatory) variables were the child's temperamental traits: general activity, sleep activity, approach, flexibility, positive mood, sleep regularity, eating regularity, habit regularity, concentration, and persistence.

The analysis included data from 402 cases, of which 301 were used as the training set and 101 as the test set. An ε-insensitive SVR regression model was applied with a capacity parameter set to 7000 and an epsilon value of 0.200. Radial Basis Function (RBF) kernels were used with the gamma parameter set to 0.100.

The model employed 182 support vectors, 160 of which were bound (support vectors with non-zero weights). The high number of support vectors indicates that many data points lay within the error margin or on its boundary, which reflects the complexity of the relationship between the predictors and the dependent variable.

The test error was 0.158, indicating relatively low deviation between predicted and actual values and suggesting good model fit. The mean squared error (MSE) was 184.864 for the training set and 142.915 for the test set. For the entire dataset, the MSE value was 174.325. Interestingly, the lower test set error may indicate good generalization and no overfitting, although the discrepancy could also point to possible underfitting.

The standard deviation ratio (STD ratio) was 0.816 for the training set, 0.878 for the test set, and 0.824 for all data. These values below 1 indicate moderate variability of model errors relative to data dispersion – suggesting model stability in prediction.

The correlation coefficients between the SVR model predictions and actual values of the dependent variable were 0.582 for the training set, 0.483 for the test set, and 0.569 for the entire dataset. This indicates a moderate predictive ability of the model – better than random guessing, but not particularly strong. The higher correlation in the training set suggests a better fit to training data than to test data.

In conclusion, the SVR model achieved moderate accuracy in predicting parenting difficulty levels based on children's temperamental traits. These results are consistent with those obtained using artificial neural networks (see Section 17.6), confirming that temperamental traits may be relatively weak predictors of parenting difficulties experienced by parents.

Comparison of the SVR Solution with Results Obtained by the ANN Model

The regression analysis using the Support Vector Regression (SVR) model and the earlier analysis conducted with an Artificial Neural Network (ANN) provided valuable insights into predicting the level of parenting difficulties based on a child's temperamental traits. Below is a comparison of the results obtained by both models. The results of the ANN model are described in detail in Section 17.6.

The ANN model struggled to accurately predict high values of the dependent variable, which translated into low predictive accuracy. It achieved an accuracy of 0.400 on the test set, indicating substantial differences between predicted and actual values. The model scored 0.574 on the training set and 0.612 on the validation set. The correlation coefficient for the ANN model was moderate, but the performance on the test set indicated issues with model fit.

The SVR model was configured with a capacity parameter of 7 and epsilon set to 0.2. A Radial Basis Function (RBF) kernel was applied with a gamma value of 0.100. The model used 182 support vectors, of which 160 were bound. The test error was 0.158, indicating small deviations between predicted and actual values of the dependent variable. The mean squared error (MSE) for the training set was 184.864, while for the test set it was 142.915, and for the entire dataset – 174.325. The standard deviation ratio was 0.816 for the training set, 0.878 for the test set, and 0.824 for all data, indicating moderate variability in the predictions. The correlation coefficients were 0.582 for the training set, 0.483 for the test set, and 0.569 for the overall data, suggesting a moderate relationship between predicted and actual values.

When comparing both models, it is evident that the SVR model achieved a lower test error (0.158) compared to the ANN model, which struggled particularly with predicting high values of the dependent variable. The prediction accuracy of the ANN on the test set was low (0.400), indicating issues with generalization.

In contrast, the SVR model showed more consistent results across the training and test sets. The correlation coefficients of 0.582 for the training set and 0.483 for the test set point to a moderate strength of association and better overall model fit. The smaller differences between sets in the SVR model suggest greater model stability and a better ability to generalize patterns from the data.

18.1.1. Application of Support Vector Machines (SVM) in Verifying Psychological Models

Modern psychology increasingly relies on advanced mathematical tools and analytical techniques to verify theoretical models. Classical circular models, widely used in psychology and education, allow for intuitive representation of relationships between variables, yet they face significant limitations due to their two-dimensional geometric structure. The problem of "dimension compression" leads to less precise interpretation, especially when the number of variables increases and their positions begin to overlap in the representational space.

In response to these difficulties, Szymańska (2025a) proposed an extension of circular models into a three-dimensional form (Szymańska, 2025a). In such models, variables are arranged in three-dimensional space, avoiding the "compression" effect and significantly improving the representation of both the intensity and direction of relationships between variables. The additional dimension enables the modeling of hierarchical structures and complex psychological relations that remain invisible in a two-dimensional layout. In practice, these models can be used to analyze theories in which the number and configuration of variables exceed the interpretative capacity of classical circular models.

To theoretically verify such models, Szymańska proposes the use of Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel as an analytical tool that allows for the examination of both classical and three-dimensional forms of the model. This method is based on the ability of SVM to transform data into higher-dimensional space — specifically, into a Hilbert space with a reproducing kernel (RKHS) — making it possible to capture nonlinear relationships between variables and more accurately reflect theoretical assumptions. Even in the case of circular models, the application of the RBF kernel allows for testing them in three--dimensional space, where data structures become clearer and relationships more pronounced (Szymańska, 2025c).

This approach has been further developed in three of the author's publications (Szymańska, 2025a, 2025c, 2025e), which demonstrate that Support Vector Machines with an RBF kernel can serve not only as tools for data classification but also for the verification of psychological models by analyzing the geometric distribution of individuals in feature space. In contrast to classical methods that operate on relationships between variables, this method enables the transfer of the model into the person space — where specific configurations of individual profiles become carriers of the theoretical structure. If the model exists, it manifests as a spatial structure — for example, a cluster of individuals whose configuration of traits aligns with a circular model. In this sense, SVM with an RBF kernel not only expands the measurement space but also allows for empirical detection of the model's presence in the population structure itself.

In other words, we no longer check merely whether "trait A correlates with trait B", but rather examine whether people with specific configurations of traits

form a geometric arrangement as predicted by the model. If a theory assumes that certain configurations of psychological traits should co-occur while others should exclude one another, then individuals representing these patterns should be located close together or far apart in the geometric space — in line with the model's assumptions. If the observed distribution of individuals in this space reflects the structure predicted by the theory, it can be considered that the model is confirmed by the empirical configuration of the data, and not only at the level of formal equations (Szymańska, 2025c).

Thus, the SVM method with an RBF kernel allows us to determine whether a given psychological model truly exists in the population — not through variable analysis, but through the way individuals are arranged in relational space. This is a new approach to model verification — where people themselves become the carriers of the theory, rather than just numbers in tables (Szymańska, 2025c).

18.2. Naive Bayes Classifier: A Simple but Effective Solution

The Naive Bayes Classifier is one of the simplest, yet most effective machine learning algorithms used in classification. It is based on Bayes' theorem, which connects conditional probability with prior probabilities, enabling the construction of models capable of predicting the class of input data (Nisbet et al., 2009). Despite its simplicity, the Naive Bayes Classifier is widely applied in practice due to its efficiency and scalability, particularly in text classification tasks such as spam filtering, sentiment analysis, or document categorization.

In this chapter, we will take a closer look at the Naive Bayes Classifier, discussing its fundamental assumptions, mathematical foundations, and examples of its application in real-world problems. We will begin by reviewing Bayes' theorem and the model's assumptions, then proceed to implementation details and an analysis of its strengths and limitations.

Bayes' Theorem and Fundamental Assumptions

Bayes' theorem is a fundamental tool in statistics and probability theory, allowing the updating of prior beliefs in light of new evidence (data). This theorem connects conditional probability with prior probabilities.

Mathematically, Bayes' theorem is expressed as:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Where:
P(A|B) is the probability of event A given that event B has occurred (known as conditional probability).

P(B|A) is the probability of event B given that event A has occurred.
P(A) is the prior probability of event A (before incorporating evidence).
P(B) is the prior probability of event B.

The Naive Bayes Classifier introduces an additional assumption that the input features are conditionally independent, meaning that the influence of one feature on the classification outcome is independent of the influence of other features. This assumption simplifies the computations and allows for an efficient application of Bayes' theorem in practice.

Example: Application of Bayes' Theorem in the Context of Parental Withdrawal from the Relationship with a Child

In this example, the probability that a parent is experiencing parental difficulties will be estimated, given that they have withdrawn from the relationship with the child. It is known that the parent has withdrawn from the relationship, is not interested in the child, and does not want to maintain contact. To precisely analyse the situation, Bayes' theorem will be applied. The details and calculations are presented below.

Data
Based on the research by Szymańska (2017d, 2019), let us assume the following (Szymańska, 2017d, 2019):

1. $P(D_{high}) = 0.25$ – The probability of experiencing high parental difficulties, estimated based on the knowledge that 25% of individuals in the population experience such difficulties.
2. $P(W) = 0.30$ – The probability of parental withdrawal from the relationship with the child, estimated based on the knowledge that 30% of individuals in the population tend to withdraw from the relationship with the child.
3. $P(W|D_{high}) = 0.78$ – The probability of withdrawal from the relationship with the child if experiencing parental difficulties. Based on Szymańska's studies, it is known that as the experience of difficulties increases, withdrawal increases sharply and amounts to approximately 78% in situations of high difficulties.

Objective
The objective is to calculate the probability that a parent is experiencing parental difficulties, knowing that they have withdrawn from the relationship with the child. Formally, we calculate $P(D_{high}|W)$.

Application of Bayes' Theorem
To calculate $P(D_{high}|W)$, Bayes' theorem is used, which is expressed by the formula:

$$P\left(T_{high} \mid W\right) = \frac{P\left(W \mid T_{high}\right) \cdot P\left(T_{high}\right)}{P\left(W\right)}$$

Substituting the known values into the formula, we obtain:

$$\mathrm{P}\left(\mathrm{T}_{\text{high}} \mid \mathrm{W}\right) = \frac{0.78 \cdot 0.25}{0.30}$$

Calculations
Let us break down the calculations step by step:

1. Calculate the product of P(W|D_{high}) and P(D_{high}):

$$0.78 \cdot 0.25 = 0.195$$

2. Then divide the result by P(W):

$$0.195/0.30 = 0.65$$

According to the calculations, the probability that a parent is experiencing parental difficulties, given that they have withdrawn from the relationship with the child, is 0.65, or 65%**.**

Summary
Based on the above calculations, we can conclude that there is a 65% probability that a parent who has withdrawn from the relationship with the child is experiencing parental difficulties. Bayes' theorem enables a better understanding of the relationships between different variables and allows for a more precise analysis of data concerning the parent–child relationship. This type of analysis is extremely valuable in research on upbringing and can contribute to the development of effective support strategies for parents.

18.2.1. Results of the Naive Bayes Classifier: An Example of Predicting Parental Difficulties

The Naive Bayes Classifier is one of the simplest and at the same time most effective machine learning algorithms used in classification. Its popularity stems from its ease of implementation, efficiency, and ability to handle large datasets. This classifier is based on Bayes' theorem and assumes that the input features are conditionally independent, which significantly simplifies computations (Nisbet et al., 2009).

In the context of classification using the Naive Bayes Classifier, it is essential to understand two fundamental types of variables: the dependent (nominal) variable and the input data (predictors).

The dependent variable is the variable we aim to predict based on other variables. It must be nominal, meaning it takes categorical values such as "yes/no", "positive/negative". In psychological data, this may include the presence of depression (0 – absence, 1 – presence) or another mental state. It is important to note that the dependent variable may have more than two levels, which allows for classification

into more than two categories, for example, different types of mental disorders (depression, anxiety, no disorder).

Input data, also referred to as predictors, may be numerical or categorical. In psychological data, predictors often include results of various psychological tests, rating scales, stress levels, or survey scores. Each predictor is treated as independent in the context of classification, which is a key assumption of the Naive Bayes Classifier.

The Naive Bayes Classifier operates by applying Bayes' theorem to estimate the probability of the input data belonging to specific classes. This process can be divided into three main stages.

The first stage is data preparation. The data must be properly prepared for analysis, which includes data cleaning, removal of missing values, normalization of numerical variables, and transformation of categorical variables into numeric form. It is also important that the training data be representative of the actual population being analysed.

The second stage is model training. The classifier learns from the provided training data by analysing the relationships between the predictors and the dependent variable. It calculates conditional probabilities for each class based on the input data. The learning process consists of estimating the model parameters, i.e. the conditional probabilities $\mathbf{P(X_i|Y)}$ for each feature $\mathbf{X_i}$ and class $\mathbf{Y}$, as well as the prior probabilities $\mathbf{P(Y)}$ for each class $\mathbf{Y}$.

The third stage is prediction. After training the model, the classifier can be used to predict the class of new data. For each new set of input data $\mathbf{X}$, the algorithm calculates the probability of belonging to each class $\mathbf{Y}$ and selects the class with the highest probability. Prediction is performed by applying Bayes' formula to calculate $\mathbf{P(Y|X)}$, where $\mathbf{X}$ is the set of input features.

The Naive Bayes Classifier is a powerful classification tool that, thanks to its simplicity and efficiency, finds broad application in data analysis. The key assumption of this algorithm is the conditional independence of input features, which simplifies the learning and prediction process (Nisbet et al., 2009). In the context of psychological data, this classifier can be used to predict various mental states based on test results and other measurable features. Prior to analysis, it is important to properly prepare the data and select appropriate variables, which ensures the accuracy and reliability of classification results.

Example – Analysis of Experiencing Parental Difficulties

In this example, an analysis was conducted to determine whether a parent is experiencing parental difficulties, taking into account two quantitative predictors: the level of discrepancy and the measure of the negative representation of the child in the parent's mind.

A binary dependent variable and two quantitative predictors were used to perform the analysis. The dependent variable, representing parental difficulties, takes the values 0 or 1, where 0 indicates no difficulties and 1 indicates the presence of difficulties.

The quantitative predictors include discrepancy, which is an indicator of the extent to which the child does not develop the psychological traits expected by the parent, and negative representation, which quantifies the parent's negative perception of the child.

The aim of the analysis was to examine whether, based on these predictors, it is possible to accurately predict whether the parent is experiencing parental difficulties.

The calculations were performed using data from the sample described in Appendix C. The variable concerning the experience of parental difficulties was standardized, and individuals with scores above and below half a standard deviation to the right and left of the mean were included in the analysis. As a result, 156 individuals from the original 319 were included in the sample. Then, during the analysis, this set of 156 individuals was divided into a training and a test set in a 75% to 25% ratio. Thus, the test set contained 39 observations, while the model was trained on the remaining 117 observations.

Table 18.2 presents the classification results for two categories: 0 (no difficulties) and 1 (experience of difficulties). The data are divided into the number of cases correctly classified as accurate and those misclassified.

Table 18.2. Classification Summary

Class	Total	Accurate	Misclassified	Accurate (%)	Misclassified (%)
0	23	20	3	86.96%	13.04%
1	16	13	3	81.25%	18.75%

In the case of class 0, representing no difficulties, a total of 23 cases were evaluated. Of these, 20 were correctly classified, which accounts for 86.96% accurate classifications. Only 3 cases were misclassified, representing 13.04% misclassified classifications.

Similarly, for class 1, indicating the experience of difficulties, 16 cases were evaluated. Among them, 13 cases were correctly classified, which translates to 81.25% accurate classifications. Three cases were misclassified, accounting for 18.75% misclassified classifications.

The results indicate high classification accuracy for both classes. For class 0, the accuracy was 86.96%, suggesting that the model effectively identifies cases of no difficulties. Similarly, for class 1, the accuracy was 81.25%, demonstrating the model's ability to correctly recognize cases of experiencing parental difficulties.

Despite the relatively low classification error rates—13.04% for class 0 and 18.75% for class 1—it is worth noting that the number of cases in each class differs. Class 0 contains 23 cases, whereas class 1 includes 16. This distribution may affect the classification results and should be taken into account when assessing the overall effectiveness of the model.

In summary, the Naive Bayes Classifier model demonstrates high effectiveness in predicting parental difficulties based on the analysed predictors. The high classification accuracy rates and relatively low error rates indicate the robustness of the model in identifying both cases of no difficulties and cases of experiencing parental difficulties.

Effectiveness Measures

Accuracy is a measure of the effectiveness of a classification model, indicating the proportion of correctly classified cases relative to the total number of cases. The formula for accuracy is as follows:

$$\text{Accuracy} = \frac{\text{Number of correct classifications}}{\text{Total number of cases}} = \frac{(20+13)}{(23+16)} = \frac{33}{39} \approx 84.62\%$$

An accuracy of 84.62% indicates that the majority of cases were correctly classified. This is a solid indicator suggesting that the model has good overall performance. However, accuracy alone is not a sufficient measure to evaluate model effectiveness, especially in the case of unevenly distributed classes.

Sensitivity, also known as the True Positive Rate, measures the model's ability to correctly identify positive cases for a given class. For class 0 (no difficulties), sensitivity is calculated as:

$$\text{Sensitivity} = \frac{\text{Correct classifications for class 0}}{\text{All actual cases of class 0}} = \frac{20}{23} \approx 86.96\%$$

A sensitivity of 86.96% for class 0 means that the model is very effective in identifying cases with no difficulties.

Similarly, the sensitivity for class 1 (experience of difficulties) is calculated as:

$$\text{Sensitivity} = \frac{\text{Correct classifications for class 1}}{\text{All actual cases of class 1}} = \frac{13}{16} \approx 81.25\%$$

Likewise, a sensitivity of 81.25% for class 1 shows that the model also performs well in detecting cases of experiencing difficulties.

Specificity, also known as the True Negative Rate, measures the model's ability to correctly identify negative cases. For class 0, specificity is calculated as:

$$\text{Sensitivity} = \frac{\text{Correct classifications as not-class 0}}{\text{All actual not-class 0 cases}} = \frac{13}{16} \approx 81.25\%$$

Specificity for class 1 is calculated as:

$$\text{Sensitivity} = \frac{\text{Correct classifications as not-class 1}}{\text{All actual not-class 1 cases}} = \frac{20}{23} \approx 86.96\%$$

Specificity for class 0 (81.25%) and class 1 (86.96%) indicates that the model effectively recognizes negative cases for both classes. High specificity means that the model rarely confuses cases of no difficulties with cases of experiencing difficulties and vice versa.

The results show that the Naive Bayes Classifier achieved high effectiveness in predicting whether a parent is experiencing parental difficulties based on the level of discrepancy and negative representation of the child. The model's accuracy of approximately 84.62% means that it correctly classifies the majority of cases. Sensitivity

and specificity for both classes are high, indicating the model's strong ability to detect both the presence and absence of difficulties.

The high values of sensitivity and specificity for both classes suggest that the model is well-balanced and effective in identifying both cases of no difficulties and cases of experiencing parental difficulties. This demonstrates the model's capacity to accurately predict outcomes based on the analysed predictors. The Naive Bayes Classifier may therefore be a useful tool in identifying parents in need of support, based on the level of discrepancy and negative representation of the child.

18.3. K-Nearest Neighbors Algorithm: When and How to Use It?

The K-Nearest Neighbors (KNN) algorithm is one of the simplest and most intuitive machine learning algorithms, used for both numerical and categorical data (Nisbet et al., 2009). This algorithm is based on the assumption that similar objects are located close to each other in the feature space. In contrast to many other algorithms, KNN does not build an explicit model during the learning process. Instead, all computations are performed at the moment of classifying a new data point. Its popularity stems from its ease of implementation and effectiveness across many different domains, such as image recognition, medical data analysis, spam filtering, and recommendation systems. The key assumption of KNN is that similar cases are located close to each other in the feature space, which enables the classification of new data points based on their nearest neighbors from the training set (Elder et al., 2012; Nisbet et al., 2009).

The KNN algorithm is particularly useful in situations where relatively small datasets are available (Nisbet et al., 2009). Its simplicity makes it an ideal tool in cases where ease of interpretation is a priority (Elder et al., 2012). However, due to its high computational complexity, KNN may be inefficient when working with very large datasets, as it requires calculating the distance between all data points.

The KNN algorithm operates by classifying new data points based on their similarity to points in the training set (Nisbet et al., 2009). KNN learns from a training dataset that contains examples with specific features and assigned labels. For instance, in a dataset concerning fruits, each fruit may be described by features such as color, size, and weight, along with a label indicating its type (e.g., apple, banana, pear, strawberry). When a new data point appears, such as a raspberry with known features but an unknown label, the KNN algorithm comes into play. The new data point is classified based on its similarity to points in the training set.

When a new data point appears, the KNN algorithm calculates the distances between this new point and all points in the training set (Nisbet et al., 2009). The most commonly used distance metric is Euclidean distance, although other metrics can be used depending on the problem's specifics. The algorithm then selects the K nearest neighbors—those points from the training set that are closest to the new point

(Nisbet et al., 2009). The choice of the K value is crucial to the algorithm's performance. A small K (e.g., K=1, K=3) makes the algorithm highly sensitive to local data patterns, which may lead to overfitting and a high susceptibility to noise in the data. A large K (e.g., K=10, K=20) makes the algorithm more stable and resistant to noise, but it may average the results, potentially leading to the loss of finer data patterns.

Among the selected K nearest neighbors, the algorithm determines which class (label) is most prevalent. The new data point is then assigned to that class (Nisbet et al., 2009). For example, if the majority of the K nearest neighbors of the new point are strawberries, the raspberry will be classified as a strawberry because it is most similar to strawberries in terms of its features.

The KNN algorithm can also be applied to numerical data, where the goal is to predict a numeric value for a new data point instead of assigning it to a specific class (Nisbet et al., 2009). The distance calculation process is the same as in the case of classification. The algorithm selects the K nearest neighbors of the new data point and then predicts the value of the new point as the average of the values of its K nearest neighbors. For example, if the three nearest fruits have weights of 150g, 160g, and 170g, the predicted weight of the new fruit is (150 + 160 + 170) / 3 = 160g.

Selecting the appropriate value of K is crucial for the effectiveness of the KNN algorithm. In practice, cross-validation is often used to find the optimal K value that provides the best model performance on the test set. A value of K that is too small may result in an overfitted model, while a value that is too large may make the model too general (underfitting).

The K-Nearest Neighbors (KNN) algorithm is an effective tool for both classification and regression that uses the existing training dataset to make decisions about new data points. The classification process involves assigning the new point to the class most frequently represented among its nearest neighbors, while in regression, the predicted value is the average of the neighbors' values. A key aspect of the KNN algorithm is choosing the appropriate number of neighbors (K), which significantly affects the quality and stability of the model.

The formal notation of the KNN algorithm can be described in several key steps.

To begin, one must define the training dataset $\boldsymbol{D}$, which consists of $\boldsymbol{n}$ data points. Each data point $\boldsymbol{x_i}$ is a feature vector, and $\boldsymbol{y_i}$ is the corresponding class label. Each data point $\boldsymbol{x_i} \in \mathbb{R}^{\mathbf{d}}$ is a vector in a ***d*-dimensional space**. The number of nearest neighbors selected for classifying a new data point is denoted as **K**.

For a new data point $\boldsymbol{x_{new}}$, the distance to all data points $\boldsymbol{x_i}$ in the training set $\boldsymbol{D}$ is calculated. The most commonly used distance metric is the Euclidean distance, which is calculated according to formula 18.1:

$$d(x_{new}, x_i) = \sqrt{\sum_{j=1}^{d} (x_{new,j} - x_{i,j})^2} \tag{18.1}$$

Among all data points in set $\boldsymbol{D}$, the $\boldsymbol{K}$ points with the smallest distance to $\boldsymbol{x_{new}}$ are selected. Let $\boldsymbol{N^K(x_{new})}$ be the set of the K nearest neighbors.

For classifying the new point $\boldsymbol{x_{new}}$, the class labels $\boldsymbol{y_i}$ of all points in the set $\boldsymbol{N^K(x_{new})}$ are collected, and the most frequent class is selected. The class $\boldsymbol{y_{new}}$ is determined as:

$$y_{new} = \arg\max_{c} \sum_{(x_i,y_i)\in N^K} x_{new} 1(y_i = c) \tag{18.2}$$

where $\mathbf{1}(\boldsymbol{y_i = c})$ is an indicator function that equals 1 if $\boldsymbol{y_i = c}$, and 0 if $\boldsymbol{y_i \neq c}$.

The KNN algorithm has many advantages. It is easy to understand and implement, does not require costly preprocessing, as all computations are performed during the classification of a new point, and is flexible—capable of being used for both classification and regression problems.

However, KNN also has its drawbacks. It can be slow and memory-intensive, especially for large datasets. The results may be highly sensitive to the value of K, which is why selecting the right K is crucial for the algorithm's effectiveness. Moreover, the algorithm is sensitive to the scale of the features, making proper scaling or normalization of the data necessary.

In summary, the K-Nearest Neighbors algorithm is a powerful tool in machine learning, particularly useful in situations that require simplicity and interpretability. However, its application requires thoughtful data analysis and appropriate parameter selection to achieve optimal results. In the following sections, we will take a closer look at example applications of KNN and practical recommendations for its implementation.

18.3.1. Results of the K-Nearest Neighbors Algorithm: An Example of Predicting the Discrepancy Between Parental Goals and the Child's Current Level of Development

This section presents the results of applying the K-Nearest Neighbors (KNN) algorithm in the context of predicting the discrepancy between parental goals and the child's current level of development. The study was conducted to understand how various variables may predict the degree of mismatch between parental expectations and the actual development of the child. The calculations were performed on a dataset derived from the sample described in Appendix C.

Several variables were used in the analysis, including one dependent variable and several input variables (predictors). The dependent variable is the discrepancy between parental goals and the child's current level of development. This variable is quantitative and measures the extent to which parental goals deviate from the actual development of the child.

The input variables include the experience of parental difficulties (*difficulty*), which measures whether the parent experiences challenges in raising the child;

negative representation of the child in the parent's mind (*representation*), which assesses how the parent perceives the child in terms of negative traits and behaviors; stress response – withdrawal (*withdrawal*), which evaluates the parent's tendency to withdraw from the relationship with the child in stressful situations; and stress response – applying pressure (*pressure*), which assesses the parent's inclination to apply pressure on the child as a stress response.

The KNN model summary provides the following information: the number of predictors is 4, the number of dependent variables is 1, the number of nearest neighbors is 5, and input standardization was applied. Averaging of values was performed uniformly. The calculations were carried out using the Euclidean distance metric to determine the distance between data points. Table 18.3 presents the regression results for the KNN algorithm.

Table 18.3. Regression results for the KNN algorithm in the context of the discrepancy between parental goals and child development

Metric	Observed Value	Predicted Value
Mean	183.59	143.20
Standard deviation	181.90	72.70
Sum of squared errors	28395.16	–
Mean error	40.39	–
Standard deviation of error	165.40	–
Mean absolute error	118.37	–
Standard deviation ratio	0.91	–
Correlation	0.42	–

The cross-validation error plot in relation to the number of nearest neighbors, shown in Figure 18.1, indicates that the lowest error was achieved at **K = 5**, suggesting that this is the optimal number of neighbors for this model. Figure 18.2 presents a comparison between observed and predicted values of the discrepancy in parental goals. The dispersion of points on the plot suggests that the model has some difficulty in accurately predicting the discrepancy values, particularly at higher values of the dependent variable.

The results suggest that the KNN model with 5 nearest neighbors is optimal in terms of minimizing cross-validation error. However, the dispersion of results on the plot of observed versus predicted values indicates certain challenges in accurately modeling the discrepancy, which suggests the need for further research and possible model modification. It is also worth noting that the mean of the predicted values was lower than the mean of the observed values, and the correlation was 0.416399697, which indicates a moderate strength of association between predicted and observed values.

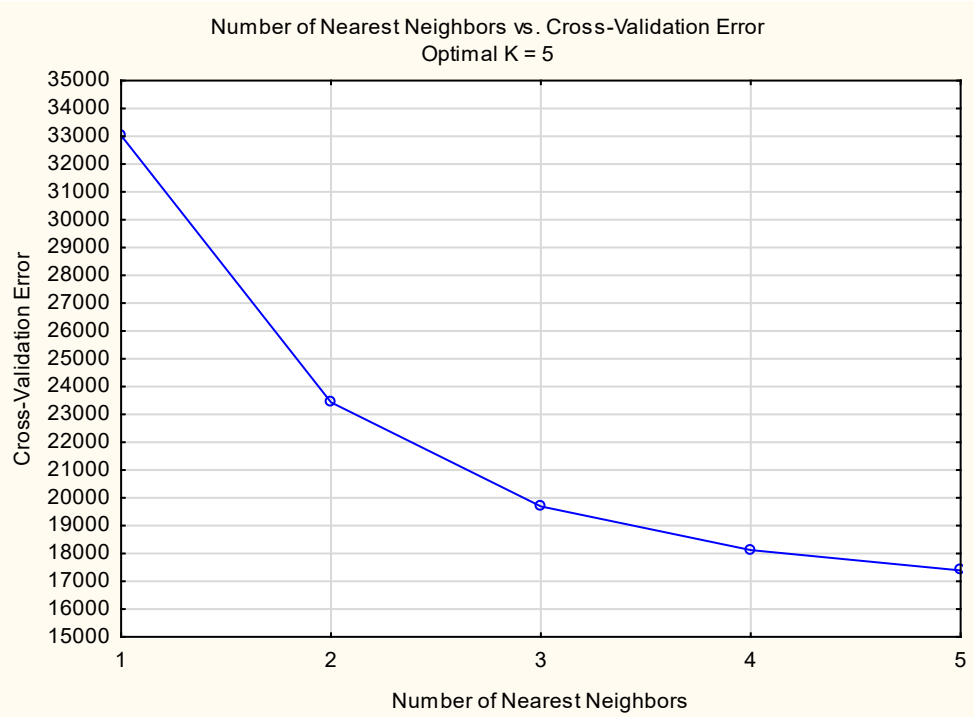


Figure 18.1. Relationship Between Cross-Validation Error and Number of Nearest Neighbors (*K*) – Optimal *K* = 5

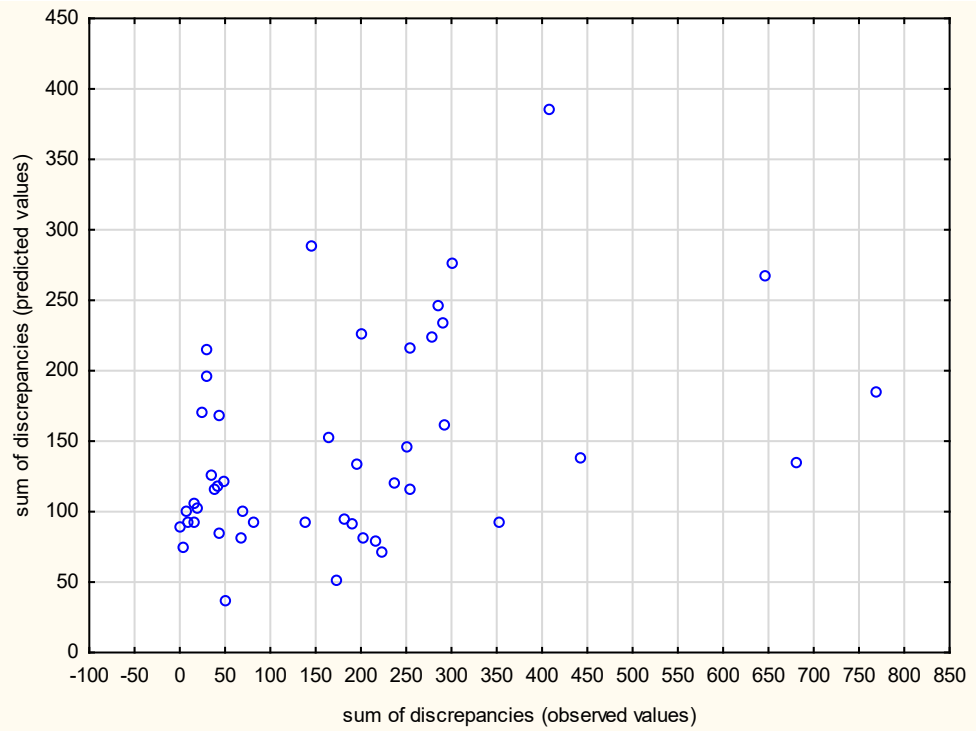


Figure 18.2. Comparison of Observed and Predicted Values of Discrepancy in Parental Goals for the Test Sample (age: 3–6 years, $K = 5$)

Interpretation of Results

The results of the KNN model indicate that with the optimal number of neighbors set at 5, the model effectively minimizes cross-validation error, as confirmed by the analyses presented in the tables and figures. Nevertheless, the model encounters difficulties in accurately predicting higher values of the dependent variable, which may stem from the algorithm's limitations in capturing more complex relationships in the data. In particular, the mean of the predicted values was lower than the mean of the observed values, suggesting a tendency to underestimate the discrepancy in parental goals.

The correlation between observed and predicted values was 0.42, indicating a moderate strength of association. While this result is promising, it is not sufficient to consider the model fully effective. It is recommended to consider using more advanced models, such as decision trees, multivariate regression, or neural networks, which may better capture the complex relationships between variables.

CHAPTER 19

Application of Machine Learning Methods in Psychological Research

Machine learning methods are a key tool in analyzing and modeling complex psychological phenomena. This chapter discusses their application in validating psychological models, with particular emphasis on artificial neural networks. These techniques enable the construction of predictive models that forecast behaviors, emotions, and other psychological aspects based on collected data.

Special attention is given to methods for assessing the operational validity of variables and the model fit of paths in structural equation modeling (SEM) frameworks. These advanced techniques allow researchers to test whether theoretical assumptions align with empirical data, which is crucial for the credibility of the research.

Later in the chapter, the use of artificial neural networks in verifying predictions based on SEM models is presented. Their application in predicting variables based on complex structural relationships is discussed. Results of the analyses are also presented, including operational and model validity, along with interpretation of training, testing, and validation outcomes.

The aim of this chapter is to present artificial neural networks as a tool that supports the validation of psychological models. These analytical techniques enable more precise research, offering researchers new perspectives for understanding human behavior and emotions. Practical examples illustrate their potential in solving research problems.

19.1. The Role of Model and Operational Validity in Verifying Relationships in Structural Models

The issue of operational and model validity in verifying relationships in structural models was the subject of a presentation delivered by Aranowska and Szymańska at a conference in Katowice in 2017. The presentation, titled *"Validity of the Latent Variable Explained by SEM Models"* (Aranowska & Szymańska, 2017), focused on methods for estimating the operational validity of a variable and the model validity of a path within structural systems. It also demonstrated how the application of artificial neural networks can support the verification of path validity in SEM models.

Structural equation modeling (SEM) is an advanced statistical method that allows both the evaluation of how well empirical models fit theoretical models and the verification of the correctness of theoretical assumptions (Bartholomew et al., 2008; Hair et al., 2006; Konarski, 2009; Szymańska, 2016b). Its key function is to assess the strength of relationships between latent variables, which are a crucial component in the analysis of interdependencies within a structural model. Within SEM models, a distinction is made between exogenous variables, which are not explained by other variables in the model, and endogenous variables, which function as both explained and explanatory variables (Hair et al., 2006; Szymańska, 2016b).

The most important variable in structural models is the main dependent variable, as it represents the endpoint of the entire analysis. All explanatory paths in the model converge upon it, making it a key element in the assessment of model validity. Empirical verification of an SEM model involves analyzing the strength of relationships between latent variables, which allows one to evaluate the extent to which these variables mutually determine each other. Strong relationships indicate that the model is well-constructed and that the explanatory variables effectively account for the variability of the main dependent variable (Aranowska & Szymańska, 2017).

When verifying a model at the empirical level, we expect to obtain information on the strength of the links between latent variables. Strong links between these variables indicate mutual dependence in the model, suggesting that they effectively explain the variability of the main dependent variable. In such cases, we also expect predictive validity, understood as the model's ability to predict the values of the dependent variable based on the explanatory variables. However, even with good model fit to the data, weak relationships between variables can limit predictive validity, making the model's results practically useless.

When relationships in the model are moderate or strong, and the model fits the data well, there is a higher probability of achieving predictive validity. Nevertheless, even in such situations, this validity may be compromised if the structural model contains incorrectly specified relationships. As Aranowska demonstrated, the validity of a structural model is undermined when latent variables explain each other more strongly than the observable variables that are supposed to better account for the variability of the latent variable they belong to (Aranowska & Szymańska,

2017). This phenomenon, also described by Hair et al. (2006), is one of the key rules for assessing model quality at the measurement level.

Although this rule has been known in the literature for many years, formulas for estimating model validity based on the principle that a variable must explain itself more than it explains other variables in the model were lacking. Our work aimed to fill this gap. Aranowska developed two formulas that are essential for assessing model validity. The first formula, known as Aranowska's gamma, is a modification of the Construct Reliability – CR formula (Szymańska & Aranowska, 2016). It was proposed that this formula be called the operational validity of a latent variable. The formal form of Aranowska's gamma was presented according to Equation 19.1:

$$\gamma = \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{\sum_{i=1}^{k} {\lambda_i}^2 + \sum_{i=1}^{k} (1-\lambda_i)^2}} = \tag{19.1}$$

$$= \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{\sum_{i=1}^{k} {\lambda_i}^2 + \sum_{i=1}^{k} 1 - \sum_{i=1}^{k} {\lambda_i}^2}} = \sqrt{\frac{\sum_{i=1}^{k} {\lambda_i}^2}{k}}$$

γ is an abbreviation for construct reliability,
i is the index of the observed variable that serves as an indicator of the latent variable,
λ is the factor loading,
δ is the error variance.

Operational validity (γ) is the square root of the sum of squared loadings (λ) of each observed variable on the given latent variable, divided by the number of these variables (k). This provides a formal tool for assessing operational validity, which is crucial in analyses based on SEM models.

The second formula, developed by Aranowska, is used to estimate **model validity** for a selected endogenous variable in the model. Model validity refers to the quality of structural relationships between latent variables, taking into account their operational validity and the strength of the paths in the model. The formal form of this formula, labeled as Equation 19.2, is presented below:

$$\nu = \sqrt{\frac{1}{p} \sum_{i=1}^{k} \sum_{j=1, j<i}^{k(i)} \gamma_j \beta_{ij}} \tag{19.2}$$

where:
i – the name or index of the latent variable,
β_{ij} – the relationship (correlation) of the i-th latent variable with the j-th one, for each preceding variable j in the network,
$k(i)$ – the number of incoming paths to the i-th variable,
p – the total number of paths leading to the i-th variable,
γ_j – the operational validity of the j-th latent variable.

Model validity (v) is the square root of the sum of products of operational validity (γ) and the path value (β) for each latent variable, divided by the total number of paths leading to that variable (p).

This formula allows for the estimation of the extent to which the latent variables in the model are consistent and strongly connected through structural relationships. Model validity accounts for both the operational validity of the variables and the strength of the mutual relationships between them.

19.2. Using Artificial Neural Networks to Verify Predictions of Models Tested with Structural Equation Modeling

This chapter presents an example of calculating operational and model validity within a structural model. Figure 19.1 shows a fragmentary structural model concerning parental mistakes. The primary exogenous variable in this model is *discrepancy*, and the three main endogenous variables are: *lack of parental control*, *aggressive directiveness*, and *constraining the child's activity*. The model also includes other endogenous variables that function both as dependent and explanatory variables. These include: *experienced parental difficulty*, *the representation of the child in the parent's mind*, *parental pressure*, and *withdrawal of the parent from the upbringing process*.

In Figure 19.1, the relationships between latent variables are illustrated by arrows, accompanied by numerical values indicating the strength of the relationships between these variables. Next to each latent variable, represented as a circle, a numerical value indicates operational validity, calculated according to Equation 19.1 (Aranowska's gamma). This value was determined based on factor loadings, i.e., the relationships between a latent variable and its observed indicators.

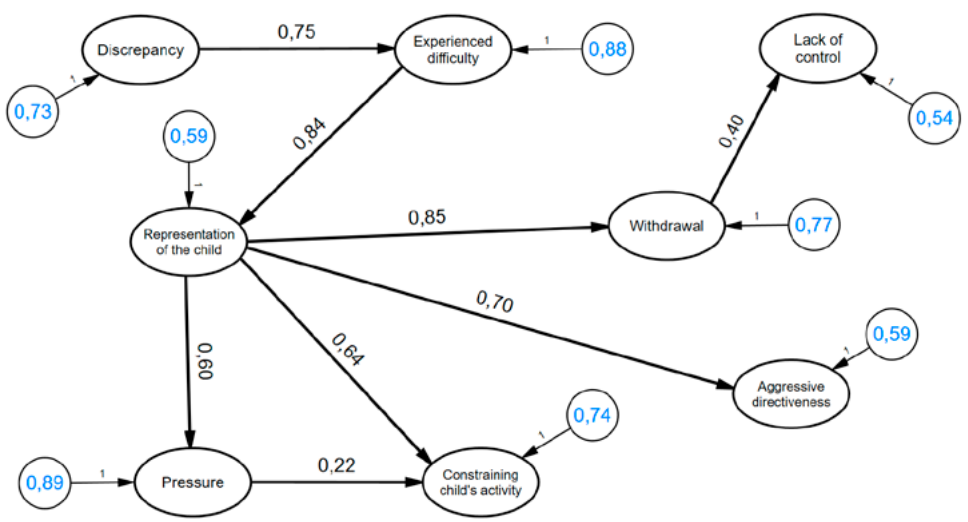


Figure 19.1. Fragmentary structural model of parental mistakes

Below, the model evaluations calculated using the method proposed by Professor Aranowska for estimating model validity are presented. The model validity for the *directiveness* path was $v^{\mathrm{D}} = 0.75$, while the operational validity of the *directiveness* variable was estimated at $\gamma^{\mathrm{D}} = 0.59$. The difference between these values is $v^{\mathrm{D}} - \gamma^{\mathrm{D}} = 0.1627$.

1. Path for directiveness

$$v_{\mathrm{D}} = \sqrt{\tfrac{1}{3}\left(0.73 \cdot 0.75 + 0.88 \cdot 0.84 + 0.59 \cdot 0.7\right)} = \sqrt{0.5666} = 0.7527$$
$$v_{\mathrm{D}} - \gamma_{\mathrm{D}} = 0.7527 - 0.59 = 0.1627$$

This means that the path explains the variable *aggressive directiveness of the parent* more strongly than the observed variables explain it themselves, which indicates a modeling error. The difference between the path value and the operational validity of the endogenous variable should be 0 or negative. Only under such conditions can the model be considered accurate. In the case of the *directiveness* path, there are serious reservations, as this path reveals an error and the variables are improperly operationalized.

Let us now examine the validity of the path related to *constraining the child's activity*.

2. Path for constraining the child's activity

$$v_{\mathrm{C}} = \sqrt{\tfrac{1}{5}\left(0.73 \cdot 0.75 + 0.88 \cdot 0.84 + 0.59 \cdot 0.64 + 0.59 \cdot 0.6 + 0.89 \cdot 0.22\right)}$$
$$= \sqrt{0.44282} = 0.665$$
$$v_{\mathrm{C}} - \gamma_{\mathrm{C}} = 0.665 - 0.74 = -0.075$$

As the calculations show, the value for the path was $v_{\mathrm{h}} = 0.665$, while the operational validity of the variable *constraining the child's activity* was estimated at $\gamma_{\mathrm{h}} = 0.74$. The difference between these values was $v_{\mathrm{h}} - \gamma_{\mathrm{h}} = -0.075$, which is close to 0 (slightly negative), indicating that the validity of the model along this path can be considered appropriate.

The last path remaining to be calculated concerns *parental control*.

3. Path for lack of parental control

$$v_{\mathrm{K}} = \sqrt{\tfrac{1}{4}\left(0.73 \cdot 0.75 + 0.88 \cdot 0.84 + 0.59 \cdot 0.85 + 0.77 \cdot 0.04\right)}$$
$$= \sqrt{0.4474} = 0.6689$$
$$v_{\mathrm{K}} - \gamma_{\mathrm{K}} = 0.6689 - 0.54 = 0.1289$$

In the case of the *lack of parental control* path, the value obtained was $\nu_K = 0.669$, while the operational validity of the control variable was estimated at $\gamma_K = 0.54$. The difference between these values, $\nu_K - \gamma_K = 0.129$, as in the case of the *directiveness* path, indicates a modeling error. The path explains the variable more strongly than the variable explains itself, which points to improper operationalization within the model.

This demonstrates that when operationalization errors occur, even strong relationships in the structural model do not guarantee the achievement of the expected predictive validity. The operational validity of variables is a necessary condition—it cannot be omitted or compensated for at later stages of analysis. If variables are poorly constructed and fail to accurately reflect their variability, it becomes impossible to obtain valid results, regardless of the quality of the analytical methods used. One cannot effectively explain a phenomenon that has been incorrectly defined at the measurement level. In the structural system under analysis, although the relationships between variables are strong and high predictive validity might be expected, operationalization errors in the variable significantly lower its level, thus preventing the achievement of reliable results.

To estimate the predictive validity of the model, we propose using an artificial neural network. This choice is based on two key reasons. First, artificial neural networks are widely recognized as among the most advanced and effective predictive models in the world. Their effectiveness in the field of prediction has been consistently acknowledged for years. Second, artificial neural networks have the ability to correct data errors during the computational procedure—an advantage emphasized by Tadeusiewicz, particularly in the case of a small number of such errors (Tadeusiewicz et al., 2007).

In other words, if the dataset contains outliers or values resulting from procedural errors, the neural network can account for these anomalies and adjust its calculations to minimize their impact on the final results. This unique ability to correct data is one of the key features of artificial neural networks, distinguishing them from other models, such as regression, which lack such capabilities.

For each path in the model, separate neural networks were constructed with the aim of predicting the value of the endogenous variable. Networks were created to explain the variables: *aggressive directiveness*, *constraining the child's activity*, and *parental control*. For each variable, 200 neural networks were constructed, from which the best-trained network—characterized by the highest predictive quality—was selected.

Prediction of Aggressive Directiveness Using Artificial Neural Networks

For aggressive directiveness, the predictors were variables connected to it within the SEM model: *discrepancy, experienced parental difficulty, and the representation of the child in the parent's mind.*

Table 19.1 presents the results of the best-trained neural network.

Table 19.1. Summary of the Best-Trained Neural Network for the Variable Aggressive Directiveness

Net-work ID	Network Name	Quality (Training)	Quality (Testing)	Quality (Validation)	Error (Training)	Error (Testing)	Error (Validation)	Training Algorithm	Error Function	Activation (Hidden)	Activation (Output)
5	MLP 3-10-1	0.456	0.695	0.75	17.044	41.051	20.897	BFGS 36	SOS	Tanh	Logistic

To estimate the predictive validity of the *aggressive directiveness* variable, an artificial neural network was employed. The best-performing network turned out to be MLP 3-10-1, which means that the model consisted of three input nodes, ten hidden neurons, and one output. The predictors in this model were the variables linked to *aggressive directiveness* in the SEM model: *discrepancy*, *experienced parental difficulty*, and *representation of the child in the parent's mind*.

The quality of the network was evaluated based on three indicators: training quality, testing quality, and validation quality. The results show that the training quality reached 0.456166, testing quality 0.695047, and validation quality 0.750659. The highest quality was achieved during validation, suggesting that the model fitted the validation data well and has potential for effective prediction on new data.

Error analysis revealed that the training error was 17.04393, the testing error 41.05150, and the validation error 20.89695. The relatively high testing error compared to the training and validation errors may indicate a generalization issue in the model. This implies that although the model fits the training and validation data well, it struggles to achieve comparable prediction accuracy on the test data.

The neural network was trained using the BFGS 36 algorithm, an advanced optimization method commonly used in neural network models. The error function used was sum of squares (SOS)**,** which is standard in this type of analysis. The hidden layer employed the Tanh activation function, allowing output values in the range from –1 to 1, while the logistic function was used at the output layer to constrain the predicted values to the range from 0 to 1.

Figure 19.2 illustrates the relationship between the dependent variable (*aggressive directiveness*) and its predicted value obtained during the training process. The horizontal axis represents the actual values of *aggressive directiveness*, while the vertical axis shows the values predicted by the model. The red line represents perfect fit, where predicted values are equal to actual values. The dispersion of points around this line suggests the model's difficulty in precisely predicting the *aggressive directiveness* variable.

The spread of points indicates a problem with the model's validity and its limited generalization capability. The poor fit confirms previous conclusions about the need to improve the operational validity of the *aggressive directiveness* variable, which is essential for enhancing the overall model performance.

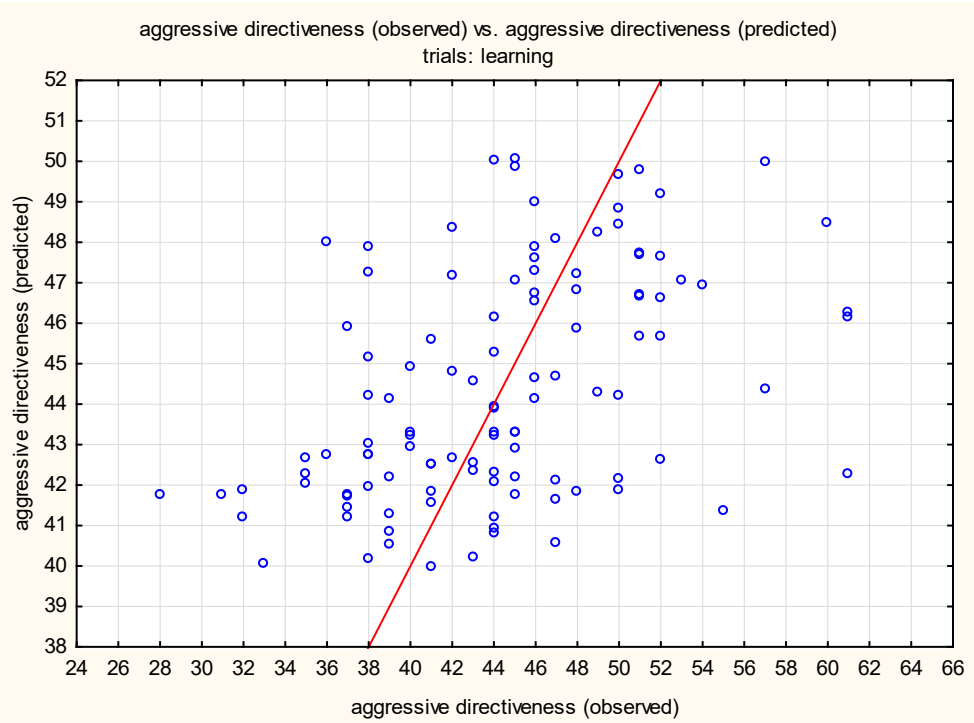


Figure 19.2. Relationship between Actual and Predicted Values of Aggressive Directiveness Estimated by the MLP 3-10-1 Neural Network

Additionally, the sensitivity analysis of the MLP 3-10-1 neural network reveals the extent to which each input variable contributes to predicting the output variable. The sensitivity values are as follows: *experienced difficulty* – 1.074250, *representation of the child* – 1.049820, and *discrepancy* – 1.022131 (Table 19.2). These results indicate that all three predictors have a relatively similar impact on the prediction of *aggressive directiveness*, with *experienced difficulty* contributing slightly more than the other variables.

Table 19.2. Sensitivity Analysis of the MLP 3-10-1 Neural Network for the Variable Aggressive Directiveness

Sensitivity analysis (doctoral_sample [new precise age] 3 to 6 years) – Training trials

Network	Experienced Difficulty (suma_tr)	Representation (reprezentacja_suma)	Discrepancy (rozbieznosc_suma)
5. MLP 3-10-1	1.074	1.05	1.022

Additional data from the prediction analysis for individual cases reveal discrepancies between the actual and predicted values of *aggressive directiveness*. These results show evident differences, confirming that the model faces challenges in prediction accuracy (Table 19.3).

Table 19.3. Predictions by the MLP 3-10-1 Neural Network for the Variable Aggressive Directiveness

Case	Discrepancy	Difficulty	Representation	Actual Aggressive Directiveness	Predicted Value	Residual
2	19	16	2	55	41.372	13.628
4	35	12	7	48	41.843	6.157
15	114	6	6	41	41.547	–0.547
16	227	11	3	45	42.185	2.815
17	223	10	9	40	42.963	–2.963
18	181	13	22	42	44.823	–2.823
19	314	30	22	52	46.617	5.383
25	25	16	14	40	43.212	–3.212
37	200	8	8	41	42.508	–1.508
38	217	2	8	35	42.034	–7.035
42	100	4	9	37	41.739	–4.739
44	56	13	8	39	42.204	–3.204
48	12	14	3	39	41.294	–2.294
49	387	8	11	44	43.904	0.097
52	220	34	16	48	45.872	2.128
54	15	21	8	38	42.737	–4.737
55	680	8	40	47	48.088	–1.088

In summary, the MLP 3-10-1 model proved to be the most effective tool for predicting aggressive directiveness among all the trained neural networks. Despite good validation results, the model requires further optimization to improve its generalization capacity and reduce errors on the test data. However, the analysis of the best-trained neural network results for the aggressive directiveness variable revealed significant problems related to model validity.

The estimated value of model validity for the path explaining aggressive directiveness was biased because the aggressive directiveness variable showed stronger associations with other variables in the model than with itself. In other words, this variable exhibited poor operational validity, which indicates the need for further work on its proper operationalization within the model.

According to the results, the testing quality (0.695) and the testing error (41.051) suggest that the model does not generalize well to test data. A possible reason for this is the weak operational validity of the aggressive directiveness variable. When a latent variable such as aggressive directiveness exhibits stronger associations with

other variables in the model than with its own observed variables, the model may struggle to accurately predict its values.

Operational validity, understood as a variable's ability to explain its own variance, is a key factor influencing the reliability of predictive results. When it is weakened, the model becomes susceptible to errors and may produce inaccurate predictions. The validation results (0.750659), along with relatively high testing errors, suggest that despite the use of a neural network, the issues related to operational validity were not fully corrected, which negatively affected the quality of prediction.

These results indicate the need for continued work on improving the operational validity of the aggressive directiveness variable. Without improving the validity of this variable, any modeling and prediction attempts may be significantly limited. Consequently, even such an advanced tool as a neural network—despite its potential—may fail to achieve high-quality prediction if the variables in the model are not properly operationalized.

Had model and operational validity not been calculated, one might have mistakenly expected high prediction quality based on strong relationships between variables in the model. However, the knowledge of biased validity already indicated potential problems with prediction. Analyses conducted using artificial neural networks confirmed these assumptions, pointing to the need for further optimization and improvement of the operational validity of variables in order to achieve more reliable and accurate predictive results.

Example of Prediction for Constraining Child's Activity

In the case of the constraining child's activity variable, better prediction results were expected, as the difference between operational and model validity was close to zero. This situation suggested that the variables used in the model were well operationalized and characterized by high validity. Therefore, there were grounds to expect that the model should be capable of more accurately predicting the values of the constraining child's activity variable.

Artificial neural networks applied in this case demonstrated potential for achieving high predictive validity due to solid theoretical foundations and well-defined variables in the model. In particular, the neural network constructed for the constraining child's activity variable had four inputs representing the following variables: discrepancy, experienced difficulty, the child's representation in the parent's mind, and applied pressure.

The following sections will discuss detailed analysis results, including the quality of prediction and model error assessment, in order to verify these expectations. Additionally, an analysis will be conducted on how the neural network performs in predicting the values of the constraining child's activity variable and whether it indeed achieves better results compared to the aggressive directiveness variable.

The MLP 4-4-1 neural network used to predict the constraining child's activity variable consisted of four input neurons, four neurons in the hidden layer, and one

output neuron. The analysis results indicated varying levels of quality in the training, testing, and validation processes. The training quality reached 0.691, the testing quality was 0.526, while the validation quality was the highest, reaching 0.812.

The similar quality levels for the training and validation sets suggest that the model generally reflected the patterns in the data well. The lower prediction quality for the test set may result from its specificity—for instance, greater case variability or lower representativeness compared to the training set. In this situation, it cannot be definitively stated that the model was unstable; rather, it is advisable to re-examine the structure of the test data and potentially apply cross-validation to confirm the model's generalizability.

The errors for the respective computation stages were as follows: 223.412 for training, 553.098 for testing, and 232.961 for validation (Table 19.4). The good quality and low validation error suggest that the model has the potential to effectively predict on new data. The relatively high error on the test set may stem from specific differences in the characteristics of the test data compared to the training and validation data, and not necessarily from problems with model generalization.

Table 19.4. Summary of Results for the MLP 4-4-1 Neural Network Summary of Active Networks

Network ID	Network Name	Quality (Training)	Quality (Testing)	Quality (Validation)	Error (Training)	Error (Testing)	Error (Validation)	Training Algorithm	Error Function	Activation (Hidden)	Activation (Output)
3	MLP 4-4-1	0.691313	0.526389	0.812273	223.4117	553.0980	232.9615	BFGS 56	SOS	Linear	Linear

The applied learning algorithm, BFGS 56 (Broyden-Fletcher-Goldfarb-Shanno), and the sum of squares (SOS) as the error function, represent a standard approach in neural networks. However, it is worth noting that this network employed a linear activation function in both the hidden and output layers, which is less typical compared to more commonly used nonlinear activation functions such as Tanh or ReLU.

The results of the sensitivity analysis indicate that the variable *pressure* was the most influential for the model's outcomes, with a value of 1.305263. The next most important variables were: *the child's representation in the parent's mind* (1.044982), *discrepancy* (1.033629), and *experienced difficulty* (1.006709). These values demonstrate that all four variables play a significant role in prediction; however, $pressure_1$ stands out as the most significant (Table 19.5).

Table 19.5. Sensitivity Analysis for the MLP 4-4-1 Neural Network

Network – Sensitivity Analysis

Network	$pressure_1$	representation_sum	discrepancy_sum	sum_tr
3. MLP 4-4-1	1.305263	1.044982	1.033629	1.006709

Table 19.6 presents the prediction results for the *constraining child's activity* variable generated by the MLP 4-4-1 neural network for various cases. Each case includes the values of the four input variables: *discrepancy* (discrepancy_sum), *experienced difficulty* (sum_tr), *the child's representation in the parent's mind* (representation_sum), and *applied pressure* (pressure$_1$). Additionally, the table contains the actual values of the dependent variable (*constraining$_1$*), the values predicted by the model (*constraining$_1$ – Output*), and the residuals, that is, the differences between the actual and predicted values.

Table 19.6. Predictions for the Constraining Child's Activity Variable by the MLP 4-4-1 Neural Network

Prediction Sheet for Constraining

Case	discrepancy_sum	sum_tr	representation_sum	pressure$_1$	constraining$_1$ (Actual)	constraining$_1$ – Output	constraining$_1$ – Residual
1	430	34	28	20	91	78.450	12.550
2	19	16	2	10	24	30.189	-6.189
4	35	12	7	4	89	22.541	66.459
5	247	34	19	10	49	49.761	-0.761
6	7	16	10	24	52	59.323	-7.323
12	104	19	32	13	51	56.418	-5.418
15	114	6	6	3	10	21.824	-11.824
16	227	11	3	13	70	41.850	28.150
17	223	10	9	4	21	29.278	-8.278
18	181	13	22	1	27	30.837	-3.837
23	0	10	2	2	19	14.678	4.322
25	25	16	14	5	53	28.705	24.295
27	19	12	7	17	64	45.010	18.991

The analysis of the results indicates that the model demonstrates varying effectiveness in fitting individual cases. In some situations, the differences between predicted and actual values are considerable. For example, in Case 4, the model predicted a value of 22.5405, while the actual value was 89.0000, resulting in a difference of 66.4595. A similar situation occurred in Case 16, where the model predicted a value of 41.8499, and the actual value was 70.0000, producing a difference of 28.1501. In Case 25, the model predicted 28.7052, while the actual value was 53.0000, yielding a difference of 24.2948.

Conversely, in some cases, the differences between predictions and actual values were small. For instance, in Case 5, the model predicted a value of 49.7605, while the actual value was 49.0000, resulting in a difference of –0.7605. In Case 1, the model predicted 78.4499, while the actual value was 91.0000, with a difference of 12.5501. Similarly, in Case 6, the model predicted 59.3232, while the actual value was 52.0000, yielding a difference of –7.3232.

These results suggest that the model may encounter difficulties in accurately predicting the values of the explained variable in certain situations, particularly when input variables take on extreme values. The large discrepancies in some cases may result from insufficient representation of such specific conditions in the training data, which indicates the need to collect a more diverse training dataset or to further optimize the model.

Figure 19.3 presents the relationship between the actual values of *constraining child's activity* and the values predicted by the model (MLP 4-4-1). The horizontal axis represents the actual values of constraining child's activity, and the vertical axis shows the values predicted by the ANN. The red line represents perfect fit, where the predicted value is equal to the actual value.

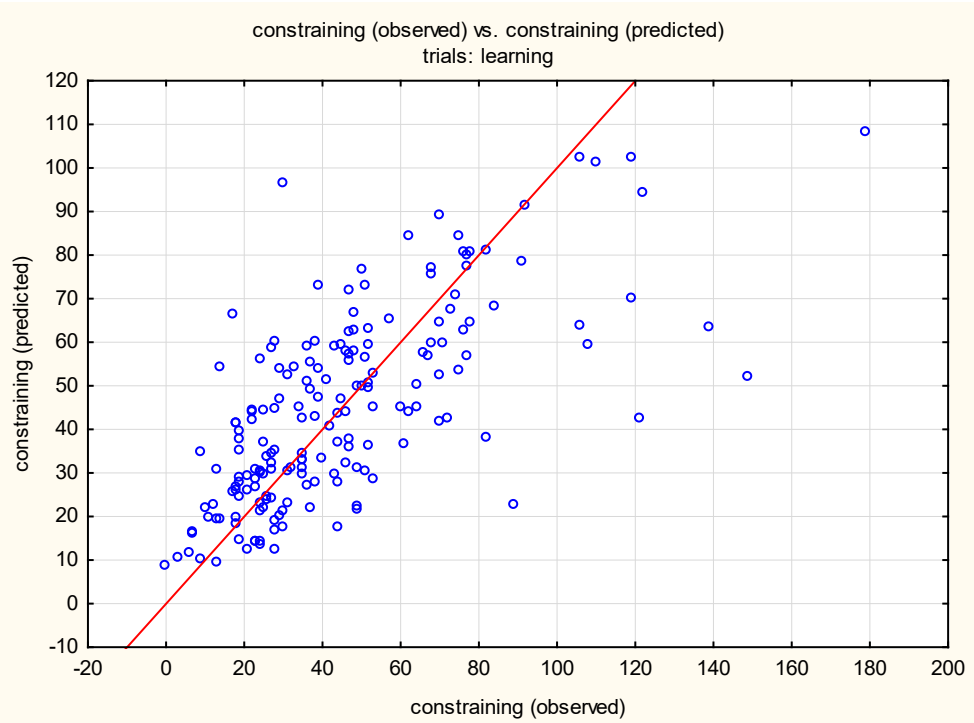


Figure 19.3. Relationship Between Actual and Predicted Values of Constraining Child's Activity by the MLP 4-4-1 Neural Network

The dispersion of points around the line of perfect fit on the plot indicates that the model experiences difficulties in accurately predicting the values of the *constraining child's activity* variable. A considerable number of outliers are visible—data points that clearly deviate from the line of perfect fit. The presence of these outliers may significantly reduce the predictive quality of the model. The large discrepancies between predicted and actual values in some cases may stem from insufficient

representation of specific conditions in the training data. This situation suggests the need to collect a more diverse training dataset or to further optimize the model.

These observations confirm the earlier conclusions drawn from the tabular analysis. They point to the necessity of continued work on improving the model to enhance its ability to more accurately predict the values of the *constraining child's activity* variable.

Comparison of Predictions for Aggressive Directiveness and Constraining Child's Activity

In the case of *aggressive directiveness*, the best prediction results were obtained using the MLP 3-10-1 network. The predictors in the model were the variables: *discrepancy*, *experienced difficulty*, and *the child's representation in the parent's mind*. The network achieved varied levels of quality: the training quality was 0.456166, testing quality reached 0.695047, and validation quality was 0.750659. The errors calculated for these processes were: 17.04393 (training), 41.05150 (testing), and 20.89695 (validation). Sensitivity analysis showed that the most influential variables in the model were *sum_tr*, *representation_sum*, and *discrepancy_sum*.

The scatter plot for *aggressive directiveness* revealed a high dispersion of points, indicating issues with the operationalization of variables. The *aggressive directiveness* variable demonstrated stronger associations with other variables in the model than with its own indicators, which points to biased operational validity. This bias negatively affected the quality of predictions, and difficulties in model generalization were clearly visible.

In the case of *constraining child's activity*, the best-fitting neural network was MLP 4-4-1. The predictors in this model included: *discrepancy*, *experienced difficulty*, *the child's representation in the parent's mind*, and *applied pressure*. The network achieved a training quality of 0.691313, a testing quality of 0.526389, and a validation quality of 0.812273. The errors calculated for these processes were: 223.4117 (training), 553.0980 (testing), and 232.9615 (validation). Sensitivity analysis indicated that the most influential variables in the model were $pressure_1$, *representation_sum*, and *discrepancy_sum*.

The scatter plot for *constraining child's activity* revealed the presence of numerous outliers, which negatively impacted prediction quality. However, the majority of data points showed small errors, suggesting that the model predicted values well for typical cases. The high validation quality indicates substantial potential for the model in predicting values of the variable on new data. Problematic were the atypical cases, which may have been underrepresented in the training set, thus hindering the network's ability to learn those specific patterns.

Moreover, the *constraining child's activity* variable did not exhibit any bias in operational validity, which contributed to the higher quality of predictions on the validation set. As a result, the model for *constraining child's activity* demonstrated better predictive validity compared to the model for *aggressive directiveness*.

The number of hidden neurons in the MLP 4-4-1 model (four neurons) is relatively small, which may indicate a less complex data structure and better operationalization of variables compared to the MLP 3-10-1 model. The MLP 4-4-1 model achieved high operational validity and better validation results, suggesting greater effectiveness in data analysis despite its simpler architecture. In contrast, the MLP 3-10-1 model, although more complex, exhibited lower prediction quality.

The comparison of prediction results for the variables *aggressive directiveness* and *constraining child's activity* revealed that the model for *constraining child's activity* achieved better validation outcomes and overall higher prediction quality. The high validation quality suggests that this model has greater potential for predicting the variable on new data, despite the presence of outliers negatively impacting overall accuracy.

In the case of *aggressive directiveness*, the main issue was the insufficient operationalization of variables, as evidenced by the considerable dispersion of points on the scatter plot. The bias in operational validity resulted in lower prediction quality and difficulties in generalizing results. In contrast, for *constraining child's activity*, the problem lay in atypical cases, which may have been insufficiently represented in the training dataset. Despite these difficulties, the MLP 4-4-1 model achieved higher validation quality and better predictive validity.

In summary, the model for *constraining child's activity* proved to be more effective in predicting the variable's values on new data compared to the model for *aggressive directiveness*. Further work on improving the quality of training data and optimizing the models may significantly enhance prediction results for both variables under analysis.

Example of Prediction for Parental Control

Low prediction quality was expected for the neural network model targeting the *parental control* variable, as the model validity was significantly impaired. The *parental control* variable was characterized by low operational validity. This situation suggests issues with the variable's operationalization, which may lead to the model's limited ability to accurately predict its values. Additionally, the *parental control* variable was very weakly associated with the preceding variable in the model structure, which also negatively impacted prediction quality and the neural network's ability to effectively generalize results.

The RBF 3-23-1 neural network selected to predict the *parental control* variable demonstrated very low quality in the training, testing, and validation processes. The training quality was 0.382338, the testing quality was 0.269977, and the validation quality reached only 0.200802. These values clearly indicate the overall low effectiveness of the model, suggesting serious issues with data fit. The lowest quality was observed during validation, which may indicate the model's difficulty in generalizing and predicting on new data.

The errors for the respective processes were as follows: 4.890373 in training, 4.532952 in testing, and 13.93277 in validation. The largest error occurred during

validation, further highlighting the model's difficulty in predicting values for the *parental control* variable. The high validation error indicates a substantial discrepancy between predicted and actual values (Table 19.7).

Table 19.7. Summary of Results for the RBF 3-23-1 Neural Network

Summary of Active Networks

Network ID	Network Name	Quality (Training)	Quality (Testing)	Quality (Validation)	Error (Training)	Error (Testing)	Error (Validation)	Training Algorithm	Error Function	Activation (Hidden)	Activation (Output)
1	RBF 3-23-1	0.382	0.270	0.201	4.890	4.533	13.933	RBFT	SOS	Gaussian	Linear

The neural network was trained using the RBFT (Radial Basis Function Training) algorithm and the sum of squares (SOS) error function. A Gaussian activation function was used in the hidden layer, while a linear activation function was applied in the output layer.

The results of the sensitivity analysis (Table 19.8) for the RBF 3-23-1 neural network indicate the extent to which the input variables contribute to the value of the output variable. The sensitivity values were: *sum_tr* – 1.203361, *representation_sum* – 1.202174, and *discrepancy_sum* – 1.151763. These results show that all three predictors contribute almost equally to the prediction of the *parental control* variable, with a slight advantage for the *sum_tr* variable. This suggests that each of these variables is relevant for the model, and their influence on the output variable is relatively balanced.

Table 19.8. Sensitivity Analysis for the RBF 3-23-1 Neural Network

Network – Sensitivity Analysis

Network	sum_tr	representation_sum	discrepancy_sum
1. RBF 3-23-1	1.203	1.202	1.152

Table 19.9 presents the prediction results for the *parental control* variable across various cases. Each case includes values of the three input variables: *discrepancy* (discrepancy_sum), *experienced difficulty* (sum_tr), and *the child's representation in the parent's mind* (representation_sum). Additionally, the table contains the actual values of the dependent variable ($control_1$), the values predicted by the model ($control_1$ – *Output*), and the residuals—that is, the differences between the actual and predicted values.

Table 19.9. Prediction Sheet for the Parental Control Variable

Prediction Sheet for Control

Case	discrepancy_sum	sum_tr	representation_sum	control$_1$ (Actual)	control$_1$ – Output	control$_1$ – Residual
2	19	16	2	22	22.038	–0.037
4	35	12	7	24	20.987	3.013
15	114	6	6	18	17.890	0.110
16	227	11	3	23	21.131	1.869
17	223	10	9	25	22.291	2.709
18	181	13	22	20	19.353	0.647
19	314	30	22	23	19.914	3.086
23	0	10	2	25	22.017	2.983
25	25	16	14	24	22.095	1.905
29	271	18	15	17	20.218	–3.218
31	498	35	2	26	21.415	4.585
32	354	80	37	23	21.501	1.498

These results indicate that the model fits some cases better than others. For example, in Case 4, where the model predicted a value of 20.987 and the actual value was 24, the difference amounted to 3.013. Similarly, in Case 31, the model predicted 21.415, while the actual value was 26, resulting in a difference of 4.584.

On the other hand, in some cases, the differences between predicted and actual values were minimal. For instance, in Case 2, the model predicted a value of 22.037, while the actual value was 22, yielding a difference of only –0.037. Similarly, in Case 15, the model predicted 17.89, and the actual value was 18, with a difference of just 0.11.

Figure 19.4 presents the relationship between the actual values of the *parental control* variable and the values predicted by the RBF 3-23-1 model. The horizontal axis represents the actual values of *parental control*, while the vertical axis shows the predicted values. The red line on the graph represents perfect fit, where the predicted values equal the actual values.

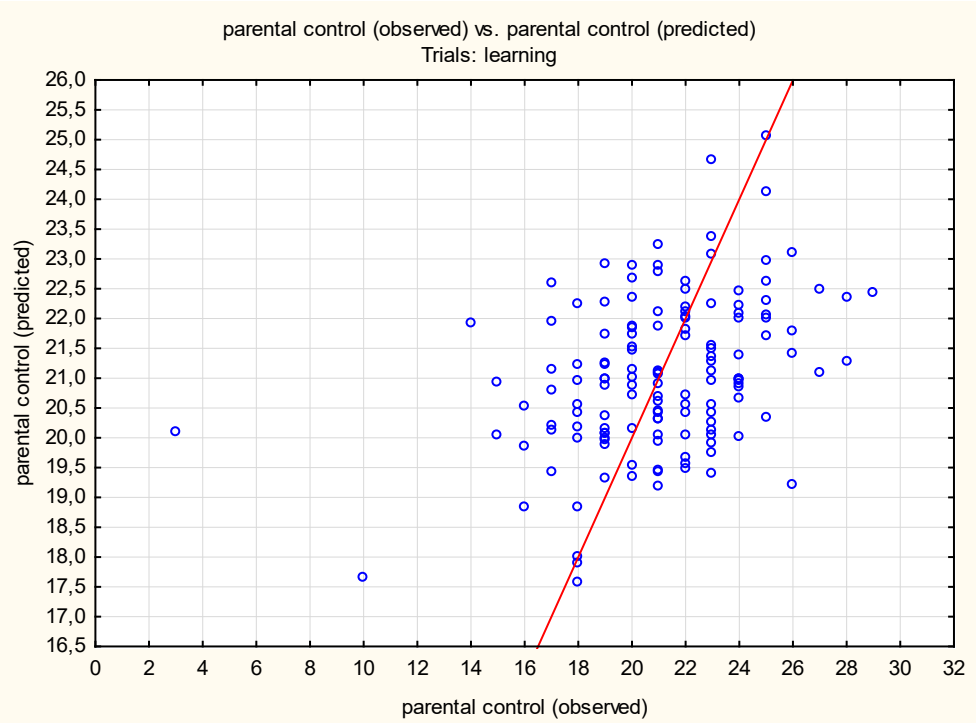


Figure 19.4. Relationship Between Actual and Predicted Values of Parental Control by the RBF 3-23-1 Neural Network

The dispersion of points around the line of perfect fit indicates the model's difficulty in accurately predicting the *parental control* variable. A visible shift of the point cloud to the right suggests the presence of systematic prediction errors. It is worth noting that in the scale of the *parental control* variable, even small differences are meaningful, which further highlights the model's weakness. The results indicate that the neural network achieved exceptionally low predictive effectiveness for this variable.

Significant discrepancies between actual and predicted values in some cases may result from insufficient representation of specific conditions in the training dataset. This situation suggests the need to collect a more diverse dataset or to further optimize the model in order to improve prediction quality.

PART V

Language Processing and Web Search Technologies

CHAPTER 20

Introduction to Speech Recognition Technologies

Artificial intelligence has existed for decades, evolving within research laboratories and academic settings. It was employed in data analysis, mathematical modeling, and expert systems, yet for most of society, it remained invisible—functioning primarily as a tool for specialists.

The breakthrough occurred in 2020, when the development of language models enabled natural, fluent communication with humans. Only then, when artificial intelligence began to converse, answer questions, translate, and offer advice, did it become part of everyday life. However, in the context of this communication, not only language models themselves are essential, but also speech recognition technologies and advanced methods of text analysis. Without them, interaction with AI would be incomplete—humans would not feel understood if artificial intelligence could not grasp meaning, intention, emotion, or context.

For this reason, this part of the book will discuss not only language models, but also other key components that enable machine understanding of content. Language models constitute an essential element of natural language processing (NLP) systems, yet their effectiveness depends on cooperation with technologies that allow meaning to be extracted from text. These technologies include, among others, text mining methods, sentiment analysis, and content classification. Without their contribution, the model would be unable to capture the message's meaning, emotion, or context, making interaction with humans superficial and devoid of depth.

Special attention will be devoted to ethical issues in the use of these technologies and to examples of their practical application, including in the analysis of

internet content, where algorithms are continuously used to scan, classify, and interpret texts—shaping users' everyday experiences online.

This part concludes with a presentation of how algorithms are used in the process of internet scanning—through web crawling and web scraping—as well as in text analysis. It will also illustrate the crucial role of these algorithms in scientific research, where they make it possible to transform massive sets of unstructured data into structured databases, forming the foundation for further analysis and scientific discovery.

Due to their fundamental importance, particular attention will be given to the BERT and GPT language models, which the author considers groundbreaking in the popularisation of artificial intelligence, making it accessible and omnipresent in daily life.

The structure of this part of the book has been designed to help readers understand the role and significance of specific language processing technologies.

This section begins with an overview of speech recognition technologies, as this constitutes the first stage in human–AI interaction—before interpreting the text, AI must first hear the utterance and convert it into written form.

Next, natural language processing (NLP) will be presented, regarded as the heart of modern artificial intelligence. NLP models enable language analysis and generation, and within this field, models based on transformer architectures—such as GPT and BERT—are of particular importance. The choice of these models is not incidental; they are the very ones that led to the popularisation of AI, making it accessible to a broad audience. They enabled not only human–machine communication but also a deeper understanding of context and emotion, transforming the perception of artificial intelligence.

Following the discussion of NLP, the technology of text mining will be introduced, as it is this process that allows information to be extracted from vast text data sets. Including text mining after NLP is a natural progression—first, the mechanisms by which AI understands language are discussed, and then it is shown how it can analyse, categorise, and extract knowledge from texts.

The next section will focus on sentiment analysis—algorithms that allow AI to recognise users' emotions and intentions. Without this form of analysis, interaction with AI would be superficial, as emotions in communication are crucial, and their recognition enables AI to better respond to human needs.

Subsequently, thematic modeling strategies and the construction of semantic networks will be discussed. These are essential for organising and interpreting large collections of textual data, enabling the formation of connections and structures within the data.

An important element of this part is also a chapter dedicated to ethical and privacy-related aspects. After presenting the tools and technologies, it becomes necessary to highlight the challenges associated with the responsible use of algorithms, particularly in the context of personal data and its protection.

This part concludes with practical workshops and techniques related to web crawling and web scraping. The order is intentional—after learning the theoretical foundations, algorithms, and their applications, the reader will be able to observe

how these technologies operate in practice, including how algorithms convert unstructured internet data into structured knowledge bases, essential for scientific research.

This structure guides the reader step by step—from the basics of speech recognition, through key language models, to practical applications and ethical challenges. Each successive section builds on the knowledge previously acquired, allowing for a deeper understanding of the importance and role of the technologies discussed.

20.1. Basics and Mechanisms of Speech Recognition

Speech recognition is an advanced field of technology that integrates elements of linguistics, computer science, and audio engineering. Its primary goal is to transform human speech into text or to initiate specific system actions in response to spoken input. Contemporary speech recognition systems rely on sophisticated statistical models and deep neural networks, which enable the automatic interpretation of acoustic signals and their translation into linguistic structure (Montejo-Ráez & Jiménez-Zafra, 2022).

The core of these systems lies in acoustic models, which describe the relationships between speech sounds and their corresponding phonetic units. The most commonly used approaches include Hidden Markov Models (HMMs) and deep neural networks trained on large datasets consisting of speech recordings and their transcriptions. This enables effective prediction of phoneme sequences and mapping of audio signals to specific linguistic elements (Blunsom, 2004).

At the same time, language modeling plays a critical role by estimating the probability of occurrence of specific word sequences. This involves both classical probabilistic models based on n-grams and modern neural models capable of capturing the context and semantic complexity of natural language. Such modeling significantly enhances the quality and accuracy of final speech recognition outcomes (Montejo-Ráez & Jiménez-Zafra, 2022).

Another key component involves alignment and decoding algorithms, which integrate acoustic and linguistic data to generate coherent text. These algorithms determine which words best correspond to the analysed speech fragment, ensuring precision and efficiency of the system's performance.

In practice, the effectiveness of speech recognition systems largely depends on their ability to reduce noise and adapt to individual speaker characteristics. Thanks to personalisation, these systems can operate with high accuracy even in challenging acoustic conditions and across diverse users—an especially crucial factor in applications that require high precision.

Finally, it is important to emphasise the role of user interfaces, which are designed to make interaction with the system intuitive and user-friendly. As a result, speech recognition is applied across many domains—from voice assistants and navigation systems to assistive technologies for individuals with disabilities—demonstrating the versatility and adaptability of this technology.

New Approaches: NLP, Deep Learning, and Language Models

Speech recognition today is no longer merely a matter of converting sound into text—it is also about understanding what was said, in what context, and with what intention. To achieve this, contemporary speech recognition systems employ advanced methods of Natural Language Processing (NLP) and deep learning techniques, which together have led to a breakthrough in this field (Montejo-Ráez & Jiménez-Zafra, 2022).

NLP is a research and technology domain concerned with how a computer can analyse, interpret, and generate human language. To make this possible, NLP uses various machine learning techniques—and in recent years, primarily deep learning, that is, methods based on deep neural networks that independently learn patterns from data.

Deep learning is therefore not a field or an application—it is a technology/method that allows for the construction of more effective models within NLP. Thanks to it, it has become possible to train complex language systems on vast corpora of textual and speech data, which has revolutionised the entire field of language recognition.

Within NLP, using deep learning, so-called language models are developed—concrete systems that learn to understand natural language. Among the best-known are GPT and BERT—models capable of analysing word meanings, recognising the context of utterances, and generating human-like responses. Language models are thus tools used in NLP, while deep learning is the technology that enables their operation.

Advanced language models have set new standards in many NLP tasks—not only in text comprehension, but also in speech recognition, machine translation, sentiment analysis, and emotion interpretation (Montejo-Ráez & Jiménez-Zafra, 2022). These models are trained on enormous datasets, which allows them to capture linguistic dependencies that earlier rule-based approaches were unable to grasp.

All of this has led to the popularisation of so-called end-to-end processing, in which the system learns from beginning to end—without the need for manual data labelling, part-of-speech tagging, or the use of rigid linguistic rules (Alharbi et al., 2021). This approach not only accelerates the speech recognition process but also makes it more flexible and natural from the user's perspective.

Modern systems employing deep learning and NLP are now capable of effectively analysing user queries and generating accurate responses—applications that include voice assistants, chatbots, and customer service systems. At the same time, their complexity makes it difficult to understand how they make decisions—which remains one of the primary challenges of contemporary technologies based on deep learning (Alharbi et al., 2021).

Applications of Speech Recognition Technologies

Modern speech recognition technologies are increasingly being applied in both everyday and specialised contexts, demonstrating their versatility and developmental potential (Esposito et al., 2021). A prominent example is interactive voice assistants, which allow users to intuitively operate devices through speech. At the same time, these systems play a significant role in assistive technologies for individuals with disabilities, enabling communication and interaction with technology by voice.

Particularly dynamic development is observed in the medical sector. Speech recognition systems support innovative diagnostic solutions, where the analysis of patients' speech can assist in identifying specific neurological or psychological disorders. Although these applications still require further research and validation, the literature highlights their growing potential in improving diagnostic accuracy and speed (Herff & Schultz, 2016).

Solutions employing natural language processing are increasingly used in the area of speech analysis, including clinical contexts. These systems make it possible to assess linguistic structure, speech dynamics, and emotional expressions, supporting diagnostic processes and organising information about a patient's mental or physical health. Contemporary approaches focus on increasing the precision of such analyses to provide useful data that assist in evaluating symptoms and the course of disorders.

Challenges

Despite rapid progress, speech recognition technologies still face significant research and technological challenges. One of the key issues is the insufficient representation of low-resource languages. Most existing models are trained on high-resource languages, such as English, which means that these systems perform less effectively on less commonly spoken languages that lack sufficient speech recordings and transcriptions (Esposito et al., 2021).

Another major challenge is the need to train models on smaller datasets and with lower computational requirements. In response to these issues, intensive research is being conducted on methods that enable more efficient use of available data, including transfer learning, which allows models trained on large corpora to be adapted to underrepresented languages.

At the same time, speech recognition systems must cope with linguistic diversity, dialects, and challenging acoustic conditions—such as background noise, interference, or variability in speaker voice. Each of these factors can significantly impact recognition quality. Therefore, continuous advancement in machine learning and artificial intelligence is essential for these systems to operate effectively across diverse contexts and environmental conditions.

Summary

Speech recognition, integrated with natural language processing technologies, remains one of the most rapidly evolving areas of artificial intelligence. Its development is driven by innovations in language modeling and deep learning methods, which enable increasingly precise and context-aware transformation of speech into text. In the coming years, particular emphasis will likely be placed on improving model efficiency, reducing the need for training data, and developing systems capable of handling multiple languages—opening new prospects for both commercial and academic applications (Alharbi et al., 2021).

At the same time, despite impressive achievements, significant challenges remain. One of them is the reliance of many systems on large training datasets, which are readily available for dominant languages such as English, but far less so for languages spoken by smaller communities—for instance, Indigenous peoples of the Americas (Chang, 2023). As a result, the development of methods that enable effective learning from limited data remains a key research direction.

Equally important are issues related to privacy and user data security—especially in the context of growing public awareness about the processing of voice information. Speech recognition systems must take these concerns into account at the very stage of technological design.

Scientific research continues to provide new solutions that enhance the quality and functionality of these systems. A literature review by Esposito and colleagues (2021) indicates that modern NLP tools can significantly support the development of more natural and comprehensible user interfaces, which is crucial for the continued advancement of intelligent interactive systems.

CHAPTER 21

Advanced Natural Language Processing (NLP) Techniques

Natural Language Processing (NLP) is an interdisciplinary field at the intersection of computer science, artificial intelligence, and linguistics, focusing on interactions between computers and human language. The primary goal of NLP is to enable machines to understand, interpret, and generate human language in a meaningful and useful way (Chopra at al., 2013).

The history of NLP dates back to the 1950s, when the first attempts at automatic translation of text between languages began. The development of this field was slow until the 1980s and 1990s, when advances in computer technologies and a better understanding of linguistics accelerated research. In recent decades, thanks to progress in machine learning and deep learning, NLP has gained prominence and is becoming increasingly advanced and effective (Ruder et al., 2019).

Text mining, often used interchangeably with the term "text analysis", is the process of extracting high-quality information from text. It employs various linguistic, statistical, and computational techniques to search through large textual datasets and extract useful data (Elder et al., 2012).

The differences between NLP and text mining can be described as follows: The first difference concerns their objective. NLP focuses on enabling computers to understand, interpret, and respond to human language in a way that is natural for humans. Text mining, on the other hand, concentrates on identifying patterns, trends, and information within large collections of textual data.

The second difference relates to scope. NLP is more concerned with understanding and generating language at the sentence and dialogue level, including tasks such as speech recognition, dialogue response generation, and machine translation. Text

mining focuses on the analysis of large text corpora and is more commonly used for data analysis, such as in documents, articles, or social media posts.

The third difference lies in identity. NLP is more technical and oriented toward understanding the mechanisms of language, while text mining is more application-driven, focusing on data extraction and analysis.

Although distinct, both fields frequently intersect and complement each other in the process of text analysis and processing. Modern NLP employs both traditional machine learning methods and contemporary deep learning approaches to understand and generate human language, opening new possibilities for applications ranging from automatic translation to interactive chatbots.

In summary, NLP (Natural Language Processing) is a field dedicated to enabling computers to understand, interpret, and generate human language. It encompasses both speech and text processing, focusing on communication between humans and machines. NLP is applied in tasks such as speech recognition, machine translation, syntactic and semantic analysis, and response generation in dialogues.

Text mining is the process of analysing large collections of textual data to identify patterns, trends, and hidden information. It focuses on the processing of written texts, with the primary aim of extracting valuable data from documents, articles, or online content—without necessarily understanding language in the context of communication.

Despite their differences, both fields often complement each other by combining natural language analysis with the exploration of large textual datasets, enabling more advanced and comprehensive research and applications.

The linguistic analysis phases in NLP include:

Morphological analysis, which concerns the structure of words and their components, called morphemes—the smallest units of meaning. At this stage, NLP decomposes words into roots, prefixes, suffixes, and endings, allowing for the understanding of their grammatical functions and their impact on sentence meaning. This is particularly crucial for highly inflected languages such as Polish or Russian, where word forms significantly influence their function within a sentence.

Syntactic analysis, which focuses on sentence structure by examining how words combine into phrases and sentences. It uses syntactic trees to represent the hierarchy and dependencies between sentence elements, which is essential for understanding who is the subject, what is the predicate, and how other elements affect the overall meaning.

Semantic analysis, which examines the meaning of language at the word and sentence level. Semantic analysis seeks to interpret the meaning of individual phrases and entire sentences, accounting for both literal and figurative meanings. This task is highly complex, as the same set of words can have different meanings depending on the context in which they are used.

Discourse integration, which investigates how individual sentences and paragraphs work together to form a coherent whole. This involves analysing how information is structured within the text—for example, how topics are developed,

arguments presented, and conclusions drawn. This phase is crucial for understanding longer and more complex texts, which cannot be interpreted sentence by sentence without losing the overall context.

The final phase, **pragmatic analysis**, explores how context influences language interpretation. This includes understanding the speaker's intent, the listener's reception, and the impact of the socio-cultural situation on communication. For instance, the same utterance may be understood differently depending on who is speaking, to whom, and under what circumstances.

These five phases of linguistic analysis illustrate how NLP tackles the challenge of understanding language in a way that mimics the human ability to process natural language. Each phase contributes a distinct element to the overall picture of how computers can interpret and generate human language in ways that are both natural and intuitive for users.

NLP finds application in a wide range of domains—from recommendation systems and web search engines, to automatic translation tools and advanced user interfaces enabling natural voice and text-based interaction. Thanks to its ability to analyse and generate natural language, NLP allows for the creation of more intuitive and efficient computer systems that can better support users in their daily tasks.

21.1. Transformer Architecture: A Revolution in NLP

The Transformer, developed by Vaswani and colleagues (2017), revolutionised natural language processing (NLP) through the self-attention mechanism, which enables the analysis of dependencies between sequence elements regardless of their distance, eliminating the limitations of earlier recurrent neural networks (RNNs) (Vaswani et al., 2017b).

The self-attention mechanism allows for dynamic assessment of the importance of individual words in context, and the parallel processing of data increases the efficiency of training and text generation. A key component of this mechanism is Scaled Dot-Product Attention, which calculates dependencies between all words in the sequence and scales the results to avoid problems associated with high values—enabling precise weighting of individual elements. The Multi-Head Attention architecture enhances the model's ability to analyse different aspects of context simultaneously through the parallel application of multiple attention mechanisms, resulting in high-quality language generation. As a result, the model can generate text effectively and precisely, taking into account both local and global context (Vaswani et al., 2017b; Yenduri et al., 2024).

In other words, the Transformer is a model that has changed the way text is analysed. Unlike older recurrent networks, which processed words one by one, the Transformer can view an entire sentence at once and determine which words are key to its meaning. The self-attention mechanism functions like an intelligent information-weighting system—the model decides which elements of the text are most relevant, regardless of their position.

To determine how strongly individual words are related to one another, the Transformer applies Scaled Dot-Product Attention. This can be compared to a process of evaluating which words in a sentence have the greatest influence on its meaning. The model calculates these relationships and scales the results to prevent numerical issues caused by large values, thereby enabling precise assignment of meaning to individual components of the text.

Yet analysing context on a single level is not enough. That is why the Transformer uses Multi-Head Attention—simultaneous application of several independent attention mechanisms. It is as if several experts were looking at the same text from different angles—one focusing on grammar, another on emotion, and yet another on sentence structure. This allows the model to better understand the text and generate more natural expressions.

Thanks to its ability to effectively model long-term dependencies, the Transformer architecture has become the foundation of modern language models such as BERT and GPT, which are revolutionising tasks involving text understanding and generation. These models are the reason why search engines better comprehend user queries, voice assistants respond more naturally, and generated texts are increasingly difficult to distinguish from those written by humans.

21.1.1. Self-Attention Mechanism: Mathematical Foundations

The self-attention mechanism constitutes one of the most essential components of contemporary language models based on the Transformer architecture. Its significance stems from its unique ability to simultaneously analyse all elements of a data sequence, which allows it to dynamically capture dependencies between words in a text. In contrast to recurrent neural networks (RNNs), which process data sequentially and often encounter difficulties in modeling long-term dependencies, the self-attention mechanism enables parallel processing of the entire sequence, significantly improving the efficiency and precision of language analysis.

Self-attention allows models to better understand the context and meaning of each word in a sentence, regardless of its position. As a result, each word can be evaluated in relation to the others, rendering the analysis more comprehensive and coherent. In practice, this means that the model is capable of focusing on those elements of the text that are crucial for understanding its content. This capability is particularly important in domains where language analysis is fundamental, such as psychology. For a psychologist, this mechanism may be compared to the process of interpreting a person's narrative, in which key words, emotions, or contexts are identified in order to comprehend the individual's internal world (Cierpka, 2004).

The aim of this chapter is to present the mathematical foundations of the self-attention mechanism, in order to demonstrate how this advanced component of language models contributes to the Transformer's effectiveness. Although its mathematical structure may appear complex, understanding it is essential for fully appreciating the potential of this technology in psychology and other social

sciences. In addition, a numerical example—presented later in the chapter—will help illustrate the operation of the self-attention mechanism for readers without a background in mathematics, facilitating its comprehension in the context of practical applications.

The self-attention mechanism is based on three key steps that enable the analysis of word dependencies in a sentence in a dynamic and parallel manner. The first step involves transforming each word in the sequence into three distinct vectors: Query (Q), Key (K), and Value (V). The Query vector reflects the query—i.e., what a given word is "seeking" in other words. The Key vector represents the identity of the word, allowing for the determination of its relevance to the remaining elements of the sequence. The Value vector contains the information that a given word contributes to the overall context. These three vectors are computed using weight matrices (W_Q, W_K, W_V), which are trained during the model's learning process. In this way, each word is mathematically represented in a form that reflects its meaning and function within the context of the sentence.

The second key stage involves calculating the self-attention weights using the Scaled Dot-Product Attention mechanism. These weights indicate how strongly a given word is associated with other words in the sentence. This process is based on computing the dot product between the Query (Q) and Key (K) vectors, which enables the model to measure the degree of similarity between individual words (Vaswani et al., 2017b). Formally, this mechanism can be expressed as:

$$\mathrm{QK}^{\mathrm{T}}$$

The result of the dot product QK^{T} determines how relevant each word is in the context of the others; however, these values may become too large, thereby hindering computational stability (Vaswani et al., 2017b). To prevent this, a scaling factor $\sqrt{d_k}$ is introduced to stabilise the computations:

$$\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_k}}$$

A softmax function then normalises the output by transforming the values into a distribution that sums to one, allowing them to be interpreted as probabilities (Vaswani et al., 2017b). In this way, the model evaluates which words are essential in the given context and to what extent they should be considered:

$$\mathrm{softmax}(x_i) = \frac{\mathrm{e}^{x_i}}{\Sigma \mathrm{e}^{x_j}}$$

The third step—also part of the Scaled Dot-Product Attention—involves weighting the values (Value, V) using the previously calculated self-attention weights. This process consists of multiplying the value V of each word by the corresponding weights, which allows information from other words in the sequence to be incorporated in proportion to their contextual significance. The final word representations

in the sentence are produced as the sum of these weighted values, whereby each word "participates" in the construction of the context, but its contribution is adjusted according to its relevance in the sentence. The entire process can be summarised by the following expression (Vaswani et al., 2017b):

$$\text{attention}(\text{Q},\text{K},\text{V}) = \text{softmax}\frac{\text{QK}^{\text{T}}}{\sqrt{d_k}} \cdot \text{V}$$

A crucial aspect of the self-attention mechanism is the scaling factor $\sqrt{d_k}$, which plays an important role in preventing computational instability. When the dimensionality (d_k) is high, the values of the dot product QK^{T} may become extremely large, potentially leading to grade instability. Scaling by the square root of d_k balances the output, enabling the model to operate in a more stable and precise manner. As a result, the self-attention mechanism can efficiently process both short and long textual sequences while maintaining a high level of contextual accuracy.

Numerical Example of the Self-Attention Mechanism: Analysis of the Sentence "The cat ran quickly"

To better understand the functioning of the self-attention mechanism, let us analyse a short example based on the sentence "The cat ran quickly". Let us assume that each word in the sentence is represented by the vectors Query (Q), Key (K), and Value (V), and that the dimensionality of the vectors (d_k) is 2. This example illustrates how the self-attention process operates step by step.

Input Data
Each word in the sentence is assigned to the matrices Q, K, and V, which are as follows:

$$\text{Q}=\begin{bmatrix}1 & 0\\ 0 & 1\\ 1 & 1\end{bmatrix}, \text{K}=\begin{bmatrix}1 & 0\\ 1 & 1\\ 0 & 1\end{bmatrix}, \text{V}=\begin{bmatrix}1 & 2\\ 0 & 3\\ 4 & 5\end{bmatrix}$$

In this notation:
The first row of each vector corresponds to the word *"cat"*,
the second to *"ran"*,
the third to *"quickly"*.

Step 1: Calculating QK^{T}
The first step involves computing the matrix product of Q and the transposed matrix K^{T}, which allows us to assess how strongly each word in the sentence is associated with the others.

The transposed matrix K^{T} is as follows:

$$\text{K}^{\text{T}}=\begin{bmatrix}1 & 1 & 0\\ 0 & 1 & 1\end{bmatrix}$$

The result of this multiplication is a 3×3 matrix as follows:

$$QK^T = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}$$

Each element in the matrix QK^T shows the degree of relationship between words in the sentence "The cat ran quickly". Higher values indicate stronger semantic association.

Interpretation Table of the QK^T Matrix

	Cat	Run	Quickly
Cat	1	1	0
Run	0	1	1
Quickly	1	2	1

Conclusions:
The value $QK^T[1,2] = 1$ means that "cat" and "ran" have a moderate connection.
The value $QK^T[2,3] = 1$ means that "ran" and "quickly" are connected.
The value $QK^T[3,2] = 2$ means that "quickly" and "ran" have the strongest connection.
The value $QK^T[1,3] = 0$ means no connection between "cat" and "quickly".

Step 2: Scaling and Applying the Softmax Function
To stabilise the results, we divide the values in the QK^T matrix by $\sqrt{d_k}$, where $d_k = 2$, so $\sqrt{d_k} = \sqrt{2}$. We obtain the following solution:

$$QK^T{}_{scaled} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 0.707 & 0.707 & 0 \\ 0 & 0.707 & 0.707 \\ 0.707 & 1.414 & 0.707 \end{bmatrix}$$

Next, we apply the softmax function to each column to transform the values into probabilities summing to 1. The formal notation of softmax is:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\Sigma e^{x_j}}$$

that is: $\begin{bmatrix} e^{0.707} & e^{0.707} & e^{0} \\ e^{0} & e^{0.707} & e^{0.707} \\ e^{0.707} & e^{1.414} & e^{0.707} \end{bmatrix}$

Substituting the values ($e^{0.707} \approx 2.028$, $e^{1.414} \approx 4.113$, $e^{0} = 1$):

$$\begin{bmatrix} 2.028 & 2.028 & 1 \\ 1 & 2.028 & 2.028 \\ 2.028 & 4.113 & 2.028 \end{bmatrix}$$

The marginal sums for the first and third columns equal: 5.056, and for the second: 8.169. All values in the matrix are divided by these sums. This gives us the final result:

$$\text{softmax}\,\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_k}} = \begin{bmatrix} 0.401 & 0.248 & 0.198 \\ 0.198 & 0.248 & 0.401 \\ 0.401 & 0.503 & 0.401 \end{bmatrix}$$

Interpretation of the Result

The output matrix shows how each word in the sentence "The cat ran quickly" attends to other words. The values in the rows represent the degree to which a given word is associated with the remaining elements of the sentence.

The first row [0.401, 0.248, 0.198] means that the word "cat" focuses mainly on itself (0.401), to a moderate degree on "ran" (0.248), and almost not at all on "quickly" (0.198). This is consistent with intuition, since "cat" is the subject of the sentence and has no direct connection to the word "quickly".

The second row [0.198, 0.248, 0.401] indicates that the word "ran" divides its attention between itself and the other words. It receives moderate attention from "cat" (0.198), but focuses much more on "quickly" (0.401), as this word describes the manner in which the movement occurs.

The third row [0.401, 0.503, 0.401] shows that "quickly" pays the most attention to "ran" (0.503), which is logical, as it functions as an adverb of manner and directly describes the verb. At the same time, it has some connection to both "cat" (0.401) and to itself (0.401), which may result from the contextual association of words in the sentence.

These results demonstrate how the self-attention mechanism in the Transformer allows the model to analyse relationships between words and dynamically assign them appropriate weights depending on their meaning in the sentence.

Step 3: Weighting the Values (V)

In the final step, we multiply the resulting softmax matrix by the matrix V in order to incorporate the information carried by the words within the sentence context. As a result, each element of the word vector is enriched with information about the other words in the sentence, but in a way that depends on the degree of attention assigned to them.

$$\text{attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\,\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{d_k}} \cdot \mathrm{V}$$

$$\text{attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \begin{bmatrix} 0.401 & 0.248 & 0.198 \\ 0.198 & 0.248 & 0.401 \\ 0.401 & 0.503 & 0.401 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 0 & 3 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 1.193 & 2.536 \\ 1.802 & 3.145 \\ 2.005 & 4.316 \end{bmatrix}$$

By multiplying the attention matrix by V, we obtain a new representation of the words in which each term is enriched with contextual information derived from its relation to the other elements of the sentence.

Thanks to this, the Transformer is able to dynamically model dependencies between words and process information much more efficiently than earlier architectures such as recurrent neural networks (RNNs).

Let us interpret this result. Each number in the Attention Output matrix reflects how the self-attention mechanism alters the representation of words in the context of the entire sentence. Each row of this matrix corresponds to a specific word in the sentence "The cat ran quickly", with the first row representing the word "cat", the second corresponding to the verb "ran", and the third representing the adverb "quickly".

Each column in the matrix describes a different vector dimension of the word, which was modified after passing through the attention layer. This means that these values are not random—they result from the relationships between words in the sentence. The first column relates to sentence structure and shows how the word functions as a part of the sentence, determining its syntactic role, for example as a subject, verb, or adverb of manner. The second column indicates how much the word depends on other words in the context of the sentence, i.e., whether its meaning changes depending on its relationships with the other sentence elements.

The interpretation of the columns derives directly from the Value matrix (V), which was used in the attention mechanism's computations. This matrix originally contained word representations in two dimensions—the first dimension reflected the sentence structure, and the second the meaning of the words in context. Since the resulting Attention Output matrix was obtained by multiplying the softmax matrix by the V matrix, the columns in the Attention Output retain the same interpretation. Ultimately, it is the Value matrix V that determines which information is transmitted and incorporated during the self-attention mechanism.

The attention mechanism allows the model to capture these dependencies and dynamically adjust the representation of words to the context of the entire sentence. As a result, words are not analysed in isolation, but their meaning is shaped by their connections with all other elements of the text.

Interpretation of the Attention Output Matrix

The Attention Output matrix reflects the new representation of words after passing through the self-attention mechanism. Each word in the sentence "The cat ran quickly" not only retains its original identity but also incorporates information from other words depending on their mutual relations in the sentence context.

Each row of the matrix corresponds to a specific word, and each column represents a different dimension of its new representation. The first column refers to sentence structure, and the second to context and word meaning.

The following table shows what information each word has absorbed and how its representation in the model has changed:

Word	Value in Column 1 (Sentence structure)	Value in Column 2 (Context and word meaning)	What does it mean?
Cat	1.193 – The cat has mainly retained its original structure as the subject.	2.536 – The cat has absorbed some information from "ran", but not from "quickly".	In its new representation, the cat is still the subject, but now it also incorporates its relationship with the verb "ran".
Ran	1.802 – "Ran" has partially absorbed information from "cat" but retains its role as a verb.	3.145 – "Ran" has strongly absorbed information from "quickly", because a verb requires an adverb of manner.	"Ran" is no longer just a verb—it now reflects that its meaning depends on the adverb "quickly".
Quickly	2.005 – "Quickly" accounts for the subject "cat", but only to a small degree.	4.316 – "Quickly" has strongly absorbed information from "ran", since it describes the manner in which the action is performed.	"Quickly" cannot exist without "ran", which is why it absorbed the most information from it.

The attention mechanism has dynamically adjusted the representation of each word, making word meanings no longer static but shaped according to context.

- "Cat" remains the subject, but to some extent incorporates the verb "ran"—in its new representation it includes not only its own identity but also information about the action it performs.
- "Ran" is no longer an isolated verb—its meaning now largely depends on the adverb "quickly". The model recognises that "ran" alone may have different meanings, but only "quickly" provides it with context.
- "Quickly" has absorbed the greatest amount of information from "ran"—because in natural language, an adverb defines the manner of action, and the attention mechanism in the Transformer automatically detected this and adjusted the vector representations of the words.

As a result, the Transformer does not analyse words separately, but models them within the context of the sentence, enabling precise understanding of the relationships between them.

Thanks to the self-attention mechanism, each word in the sentence receives a representation that takes into account its meaning in relation to the other words. Despite its mathematical complexity, this process can be compared to the natural mechanism of selective attention—the model identifies and amplifies significant dependencies while ignoring less relevant elements. This is crucial both in language analysis and in psychology, where context plays a fundamental role in interpreting speech.

The self-attention mechanism allows for dynamic and parallel analysis of relationships between words, which significantly increases the efficiency of language processing. Mathematical operations such as dot product, scaling, and the softmax function enable the model to "focus attention" on key elements of the text. In the context of psychology, this mechanism can be applied to the analysis of a patient's narrative, helping to identify recurring language patterns, emotionally significant expressions, or subtle shifts in the way experiences are described.

21.1.2. Multi-Head Attention: Parallelism in Information Processing

The Multi-Head Attention (MHA) mechanism is an advanced extension of the self-attention concept and plays a key role in the Transformer architecture. The self-attention mechanism allows for flexible assignment of meaning to words in context, but its single version may limit the model to one perspective. Multi-Head Attention addresses this problem by enabling the simultaneous analysis of text from multiple points of view.

Its uniqueness lies in the use of multiple independent attention "heads", each of which analyses the data from a different perspective. For a psychologist, this mechanism can be compared to conducting a conversation with a patient in which the therapist simultaneously follows emotions, narrative context, and subtle nuances of non-verbal communication (Grzesiuk, 2005). In this way, MHA enables the capture of various aspects of the text, such as syntactic, semantic, or structural dependencies, which significantly increases the precision and depth of analysis.

Multi-Head Attention operates in parallel, which allows for the simultaneous processing of multiple parts of the text, avoiding the limitations associated with the sequential nature of recurrent neural networks. Each attention "head" is essentially a separate self-attention mechanism that focuses on a different feature of the data. As a result, MHA can simultaneously analyse both local relations between words in a sentence and global dependencies encompassing the broader context of the utterance. Similarly to psychotherapy, where different approaches can shed new light on the complexity of a problem, MHA allows the language model to explore the multidimensional context of the text. This ability is particularly important in the analysis of complex linguistic structures, where multiple levels of meaning intertwine.

The introduction of the Multi-Head Attention mechanism into the Transformer architecture has significantly improved the ability of language models to process long and complex text sequences. Thanks to parallelism in data processing, MHA not only increases computational efficiency but also enables more precise capture of linguistic nuances. In the following sections of this chapter, we will examine the detailed structure of the MHA mechanism and its practical applications.

21.1.3. The Structure of the Multi-Head Attention Mechanism

The Multi-Head Attention (MHA) mechanism is based on distributing the workload among several independent attention "heads" that operate in parallel to process different aspects of the input data. Thanks to this structure, MHA is capable of simultaneously analysing various relationships in the text, which significantly enhances the model's ability to capture complex linguistic patterns. The details of its construction and operation are presented below.

The first step in the MHA structure involves dividing the input data into smaller subspaces. The input data, typically represented as vectors of dimension $d_{mo}d_{el}$,

are split among h attention heads. Each head handles the analysis of one of these subspaces, which has the dimensionality $d_k = d_{mo}d_{el} / h$. This division allows each head to focus on a different aspect of the data, which is crucial for comprehensive analysis. In practice, this means that different heads can attend to various linguistic features, such as syntactic, semantic, or contextual dependencies.

Each attention head then operates in parallel, applying the self-attention mechanism to its assigned subspace of data. This process involves generating Query (Q), Key (K), and Value (V) vectors for the input data and calculating attention weights, which indicate the relative importance of particular elements in the text. At this stage, each head functions independently, allowing for a multidimensional interpretation of the input data.

Once each head has completed its analysis, the resulting representations are concatenated into a single vector along the feature dimension. The resulting vector then passes through a linear transformation layer with the weight matrix W_o, which transforms the combined outputs into the final result of the Multi-Head Attention mechanism. This step enables the model to capture various contextual perspectives by integrating information from multiple attention heads into a coherent representation.

The Multi-Head Attention mechanism can be described mathematically by the following equation:

$$\mathrm{MHA(Q,K,V)} = \mathrm{Concat(head_1,head_2,\ldots,head_h)W_o},$$

where each attention head ($head_i$) operates according to the equation:

$$\mathrm{head_i = Attention(QW_{Qi},KW_{Ki},VW_{Vi})}$$

Each attention head transforms the input matrices Q, K, and V using its own weight matrices $W\_\{Q_i\}$, $W\_\{K_i\}$, and $W\_\{V_i\}$. As a result, each head performs an independent analysis, focusing on different aspects of the input data. The concatenation of all head outputs into a single vector, followed by transformation via the W_o matrix, allows the model to gain a more complete insight into the relationships within the sequence.

This is a formal description of the mechanism that explains its operation at the mathematical level. However, one can also view this process in a more intuitive way.

The complex structure of the Multi-Head Attention (MHA) mechanism enables the Transformer to process information simultaneously on multiple levels, which makes it exceptionally effective in analysing complex textual sequences. In the context of psychology, this can be compared to a situation in which a therapist simultaneously analyses the patient's emotions, the context of their utterances, and the meaning of their narrative, thus gaining a more complete and multidimensional picture of the patient's experiences.

The Multi-Head Attention mechanism allows for the simultaneous analysis of multiple aspects of the text, enabling Transformer-based models to better comprehend the complexity of language. In the sentence "Kasia found a dog that was running in the park", each attention head may focus on a different type of dependency. For example, one head may analyse the syntactic structure, identifying the relationship between "Kasia" and "found". Another head may examine the semantic meaning of words, such as the connection between "dog" and "was running", taking into account the context in which these words appear. Yet another head may deal with global relationships in the sentence, such as the connection between "Kasia" and "park", considering the overall message.

The Multi-Head Attention mechanism offers immense possibilities for application, particularly in the context of narrative analysis. In psychology, it may assist in understanding patients' utterances by identifying key emotions, themes, or structural changes. Owing to the parallel operation of different heads, the model can focus on many aspects simultaneously, providing more comprehensive analyses and supporting more effective approaches to language interpretation.

21.1.4. Vanishing Gradient Problems in RNNs vs. Transformer Advantages

Traditional Recurrent Neural Networks (RNNs) were designed to analyse sequential data, such as text or time signals, by iteratively processing the input information over time (Tadeusiewicz et al., 2007). A key element of their operation is the propagation of information through a loop, which allows the models to "remember" earlier steps of the sequence. However, with deeper data processing and an increasing number of time steps, RNNs begin to encounter fundamental problems known as vanishing and exploding grades. To understand why these problems are so critical, it is necessary to examine the notion of the grade and its role in the process of machine learning.

A grade in the context of machine learning is a vector that indicates the direction and rate of change of the cost function (also called the error function) in the parameter space of the model, such as the weights of a neural network. The cost function is a mathematical measure of the model's error during prediction. During neural network training, grades are computed through the process of backpropagation, and their value determines how the network's weights should be updated to reduce the error.

In recurrent networks, grades are especially important because they must be propagated through many time steps in a process known as Backpropagation Through Time (BPTT). At each time step, the error changes resulting from earlier data are computed, which then adjusts the model's parameters. This process allows RNNs to capture dependencies between distant points in a sequence.

In theory, RNNs should be able to learn long-term dependencies by propagating grades over time. In practice, however, grades may behave in ways that hinder effective learning. This occurs when the values of the grades begin to:

Vanish (vanishing grades): Grades become very small (close to zero). As a result, the network weights are updated minimally or not at all. This problem prevents the network from learning long-term dependencies because the influence of earlier data in the sequence is "lost" over subsequent time steps.

Explode (exploding grades): Grades take on very large values, leading to instability in the model. The neural network weights are then updated in a chaotic manner, which may prevent the model from converging to optimal solutions.

The source of these problems lies in the mathematical operations performed during grade propagation. In particular, during Backpropagation Through Time (BPTT), grades are repeatedly multiplied by the network's weight matrix, which leads to their exponential growth or decay. When the weight matrix contains values close to zero or significantly greater than one, the grades become unstable.

Vanishing grades limit the ability of RNNs to learn long-term dependencies, which means that information about earlier events in the sequence has little influence on the predictions. On the other hand, exploding grades make the training process unstable, often leading to excessively large error values and the interruption of learning.

These limitations were among the main reasons for seeking alternative architectures capable of efficiently processing long sequences of data while avoiding grade-related problems. Transformers, by introducing the self-attention mechanism, proved to be a breakthrough solution in this regard (Vaswani et al., 2017b).

The phenomenon of vanishing grades arises from the mathematical nature of error propagation in deep neural networks, particularly in recurrent networks. During Backpropagation Through Time (BPTT), grades are calculated at each time step, and their value is repeatedly multiplied by the weight matrix. If the grade value at each step is less than 1, e.g., 0.90, then repeated multiplication of these values leads to the exponential decrease of grades.

This process can be described mathematically by the equation:

$$g_n = g_0 \cdot (0.9)^n$$

where g_n is the grade after n steps, and g_0 is the initial grade. Let us analyse two cases:

- For $n = 10$ (10 steps back):

$$g_n = g_0 \cdot (0.9)^{10} \approx g_0 \cdot 0.3487$$

- For $n = 50$ (50 steps back):

$$g_n = g_0 \cdot (0.9)^{50} \approx g_0 \cdot 0.00515$$

As we can see, after 50 steps the grade becomes practically zero. This means that the weights in the initial layers of the network are not effectively updated, leading to the "forgetting" of information in long data sequences. In practice, a recurrent network is unable to capture long-term dependencies, which significantly limits its applicability in tasks requiring context analysis over many time steps.

The opposite problem to vanishing grades is the phenomenon of exploding grades, which occurs when the grade value is greater than 1, e.g., 1.1. In such cases, repeated multiplication leads to exponential growth in the grade values, which causes instability during learning.

The equation describing this process is:

$$g_n = g_0 \cdot (1.1)^n$$

Let us also analyse two scenarios:

- For *n = 10*:

$$g_n = g_0 \cdot (1.1)^{10} \approx g_0 \cdot 2.5937$$

- For *n = 50*:

$$g_n = g_0 \cdot (1.1)^{50} \approx g_0 \cdot 117.39$$

In this case, the grade values grow exponentially, leading to abrupt and uncontrolled changes in the weights. The training process becomes unstable because the model makes steps that are too large when updating weights, often resulting in the model "diverging", i.e., being unable to find an optimal solution.

Both vanishing and exploding grades are a direct consequence of the mathematical structure of backpropagation in deep neural networks. These problems were particularly significant in traditional recurrent networks and limited their ability to learn complex and long-term patterns in data. The necessity to solve these problems became one of the main motivations for developing the Transformer architecture, which—thanks to the self-attention mechanism and parallel data processing—avoids the limitations associated with grade propagation through time.

The convergence of grades in recurrent neural networks (RNNs) can be modelled using the formalism of sequence and limit theory in mathematical analysis, which enables a precise interpretation of the signal propagation dynamics over time. Grades, which are used to update the network's weights during the learning process, are propagated backward through many time steps. In this process, each grade computation involves multiplication by a value dependent on the derivative of the activation function and the network's structure. This leads to the grades gradually approaching limiting values—zero or infinity—depending on the initial values and the aforementioned factors.

When grades take values less than one, e.g., $g_n = g_0 \cdot (0.9)^n$, repeated multiplication by a number less than one at each step leads to a decreasing sequence of grades. Mathematically, this can be expressed as:

$$\lim_{n\to\infty} g_n = \lim_{n\to\infty} g_0 \cdot (0.9)^n = 0$$

As time progresses, the grades approach zero, causing the earlier layers of the recurrent network to virtually stop learning. This means that information from earlier time steps is almost completely forgotten, and the network is unable to effectively model long-term dependencies in the data.

Conversely, when grades are greater than one, e.g., $g_n = g_0 \cdot (1.1)^n$, the opposite effect is observed: the grade values grow exponentially with each time step. In this case, the mathematical expression for the limit is as follows:

$$\lim_{n\to\infty} g_n = \lim_{n\to\infty} g_0 \cdot (1.1)^n = \infty$$

The grade values quickly become enormous, which leads to weight explosion in the neural network. Such behaviour causes instability in the learning process, where weight updates are too aggressive and the network is unable to converge to optimal solutions.

Both situations—vanishing and exploding grades—share a common mathematical foundation: the grade value is multiplied by a number which, depending on whether it is less than or greater than one, causes the sequence to converge either to zero or to infinity. In practice, such behaviour limits the capacity of recurrent networks to learn, particularly in tasks that require modeling long-term dependencies in data sequences.

Understanding this mathematical mechanism enabled researchers to develop architectures such as Transformers, which eliminate these problems through parallel data processing and the self-attention mechanism. In this way, issues related to grade propagation through time no longer pose a limitation in modern machine learning models.

21.2. Transformer-Based Language Models

Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is an advanced model for natural language processing that utilises the Transformer architecture introduced by Vaswani et al. in 2017 (Vaswani et al., 2017a). A key feature of BERT is its ability to process text bidirectionally, which means the model analyses the context from both the left and right sides of each word in a sentence. This contrasts significantly with the Generative Pretrained Transformer (GPT) model, which processes text in

a single direction. The BERT architecture consists of multiple Transformer encoder layers, each of which uses attention mechanisms to efficiently model dependencies and contextual relationships in the input data (Rogers at al., 2020).

BERT is trained through a two-stage process, in which the first stage involves pre-training on large text corpora using two main tasks: the Masked Language Model (MLM) and the Next Sentence Prediction (NSP). In the MLM task, randomly selected words in a sentence are masked (hidden), and the model's objective is to predict these words based on the remaining context. The NSP task involves predicting whether two given sentences follow each other in a text. These methods enable BERT to learn rich language representations, which are then fine-tuned for specific NLP tasks such as text classification or question answering (Rogers et al., 2020).

BERT is capable of encoding complex semantic and syntactic knowledge, as evidenced by its ability to understand context, irony, and subtle word meanings in different linguistic environments. Nevertheless, studies show that the model also has limitations, for example in its understanding of negation or more complex logical structures. This highlights the need for further research into attention mechanisms and their influence on the interpretive capabilities of language models (Rogers et al., 2020).

Thanks to its flexible architecture, BERT allows for various modifications and optimisations, including techniques for reducing model size and improving computational efficiency. Examples include quantisation, knowledge distillation, and pruning, which enable the use of BERT in more constrained environments such as mobile devices or web applications. Through these innovations, BERT opens new opportunities for NLP applications while offering the efficiency required in practical deployments (Rogers et al., 2020).

BERT is one of the most influential models in the field of NLP, offering deep understanding of natural language. Its bidirectional structure and flexibility in training make it an extremely powerful tool, applicable in a wide range of domains—from automatic translation to recommendation systems and sentiment analysis. Despite its many strengths, this model remains the subject of ongoing research aimed at further improving its effectiveness and efficiency.

21.2.1. Variants of the BERT Model and Their Applications

Following the success of the BERT model (Bidirectional Encoder Representations from Transformers), researchers began developing its subsequent versions, adapting the architecture to different applications, languages, and computational constraints. The aim of these modifications was to reduce the number of parameters, improve performance, and increase the model's effectiveness across various NLP tasks.

To enable the model to operate on devices with limited computational resources, DistilBERT was introduced (Sanh et al., 2020). It uses knowledge distillation to reduce the number of parameters by 40% while retaining 97% of BERT's performance. In turn, ALBERT (Lan et al., 2020) optimises the model structure by

decomposing embedding matrices and sharing weights across layers, which reduces memory usage and improves computational efficiency. To enhance performance in language understanding tasks, RoBERTa was developed; it removes the Next Sentence Prediction (NSP) task, uses longer training sequences, and is trained on larger datasets, which improves the quality of generated representations.

To adapt the model to multiple languages, mBERT (Multilingual BERT) was created. It was trained on over 100 languages without explicit language encoding, enabling universal applicability. In addition, language-specific models have been developed, such as CamemBERT (Martin et al., 2020) for French and BETO (Cañete et al., 2023) for Spanish, which achieve better performance in national text analysis than the standard mBERT. These variants expand the capabilities of BERT, making it a versatile tool in the field of natural language processing.

21.2.2. Application of the BERT Model in Psychological Research

BERT (Bidirectional Encoder Representations from Transformers), as a language model based on Transformer architecture, introduces a new standard in text analysis by enabling the inclusion of bidirectional sentence context and precise modeling of relationships between words across a wide range of tasks.

Its ability to process language bidirectionally means that each word in a text is interpreted not only in the context of preceding words but also in relation to those that follow. This represents a fundamental shift compared to earlier sequential models, which processed text in a single direction and therefore often failed to capture important semantic dependencies.

One consequence of this architecture is the ability to more accurately capture semantic relationships in text. BERT not only recognises individual words but also their meaning in a given context, which enables the analysis of linguistic phenomena requiring interpretive processing—such as homonymy, polysemy, irony, or subtle differences in the use of synonyms.

In contrast to traditional text analysis methods based on statistical frequency models, BERT uses attention mechanisms to identify semantic connections in text regardless of their distance. As a result, it enables the analysis of more complex linguistic structures and can be applied in situations where it is important to track changes in speech patterns or identify hidden linguistic patterns.

Structurally, the model is trained using the Masked Language Model task, where some words in the text are masked and the model learns to predict them based on the remaining context. This training approach develops its ability to recognise semantic relationships between words even in incomplete or ambiguous texts, which is particularly relevant in the qualitative analysis of speech.

Additionally, BERT allows for a more abstract approach to text analysis than classical methods based on statistical word representations. This opens up possibilities for the automatic recognition of narrative and linguistic patterns that are difficult to capture directly through quantitative methods.

Thanks to its bidirectional attention mechanism, BERT is a key tool in the field of text mining, offering the ability to extract information from texts in a context-aware and semantically precise manner. Its ability to perform qualitative content analysis enables the interpretation of context-dependent meanings, which is invaluable in research on discourse structure, changes in narrative style, and hidden linguistic schemata. In psychology, it can be used for narrative analysis, exploration of language patterns, and identification of subtle differences in the way statements are formulated by different respondent groups.

Importantly, BERT is not limited to classifying texts; it enables their dynamic interpretation, which is crucial in studies involving evolving speech patterns. Its capacity to track semantic relationships between words allows for the analysis of emotional language, exploration of expression styles, and identification of hidden meanings that may escape traditional quantitative methods. As such, it can be applied in clinical, upbringing, and social psychology, where language analysis is a key component of diagnostics and research into cognitive processes. Its flexibility and ability to process large textual datasets make BERT one of the most advanced tools in psychology based on textual data exploration.

21.3. Generative Pre-trained Transformer (GPT): Unidirectional Language Modeling

This subsection discusses the architecture and functioning of the Generative Pre--trained Transformer (GPT) model based on a literature review, particularly the comprehensive analysis conducted by Yenduri et al. (2024). The Generative Pre--trained Transformer (GPT) is an advanced natural language processing model that utilises the Transformer architecture developed by Vaswani et al. (2017). In contrast to the BERT model, which processes text bidirectionally, GPT operates autoregressively, analysing sequences in only one direction—from left to right. This structure makes GPT particularly effective in text generation and language modeling in a manner that resembles human speech production.

Like other Transformer-based models, GPT uses self-attention mechanisms. However, a key difference from BERT lies in the use of masked self-attention, which blocks access to future tokens in the sequence. As a result, the model learns to predict subsequent words based on the preceding context, making it especially effective in generating coherent and logical utterances.

The name GPT—Generative Pre-trained Transformer—reflects the key aspects of how this language model operates. The term *Generative* indicates that the model not only analyses text but also generates new content by predicting subsequent words autoregressively, i.e., based solely on previously encountered tokens. Thanks to this generative nature, the model performs well in tasks requiring the production of coherent utterances, such as dialogue simulation, paraphrasing, or automated content generation.

The *Pre-trained* component refers to the two-stage training process, which includes pretraining the model on large-scale text corpora, followed by fine-tuning for specific tasks such as text summarisation or question answering. The final component, *Transformer*, refers to the architecture based on the self-attention mechanism, which enables the model to analyse dependencies between words regardless of their position in the text. This solution, introduced by Vaswani et al. (2017), marked a breakthrough in natural language processing by eliminating the limitations of earlier sequential models, such as recurrent neural networks (RNNs), which struggled with modeling long-term dependencies.

Despite its advanced architecture and broad applicability, the GPT model has significant limitations stemming from both its structure and mode of operation. Most notably, it does not understand language in a human-like way—it does not analyse semantics in a deep, conceptual sense but rather operates on predictions. This means it generates text based on the statistical likelihood of word sequences. This limitation leads to a further challenge: the potential to generate false information. The model lacks an inherent fact-checking mechanism, which makes it capable of producing plausible-sounding yet incorrect or entirely fabricated content, known as "language hallucinations". This phenomenon has been thoroughly examined in recent studies comparing GPT-3.5 and GPT-4, which show that although there have been noticeable improvements in newer versions, hallucinations remain a serious issue (Mohammed et al., 2024).

Another important challenge is the necessity of fine-tuning the model for specific tasks. Although GPT demonstrates high text-generation capabilities, its base version often requires task-specific fine-tuning to meet user needs and produce content that aligns with expectations. Without this adaptation, the model may deliver low-quality or contextually inappropriate responses. Additionally, GPT models—especially more recent versions—have high computational demands during both training and text generation. These requirements include significant hardware resources and energy consumption, which limits accessibility for smaller research institutions and businesses (Cottier et al., 2025).

21.3.1. Evolution and Variants of GPT Models

Generative Pre-trained Transformer (GPT) models have undergone significant evolution since the publication of the first version by OpenAI in 2018. GPT-1 introduced the concept of generative pretraining, which enabled more efficient model training on large datasets without the need for manual labelling. The subsequent version, GPT-2, increased the number of parameters to 1.5 billion, significantly improving the quality of generated text and the model's ability to solve a wide range of NLP tasks without fine-tuning. The greatest breakthrough came with the release of GPT-3 in 2020, which reached 175 billion parameters and introduced few-shot and zero-shot learning approaches. These approaches allowed the model to perform language tasks without prior fine-tuning on specific datasets (Brown et al., 2020).

GPT-4, released in 2023, improved contextual coherence and the ability to model complex semantic relationships, while also expanding its capabilities to include multimodality, enabling the processing of both text and images. The latest version, GPT-4o, introduced in 2024, further enhanced model efficiency by increasing accuracy and reducing language hallucinations. It also extended the model's ability to process audio in real time. Each successive iteration of the model contributes to increasing its adaptability and precision in content generation, which finds application in numerous areas—from language analysis to user interaction (OpenAI, Achiam et al., 2024; OpenAI, Hurst et al., 2024).

21.3.2. The Use of the GPT Model in Psychology

The development of models based on the Generative Pre-trained Transformer (GPT) architecture opens new possibilities in psychology, particularly in the domain of education and therapist training. Thanks to its ability to generate text in a fluent and contextually coherent manner, this model can simulate realistic interactions, provide instructional materials, and support the analysis of therapeutic communication. The use of GPT in psychology goes beyond traditional teaching methods by enabling dynamic content adaptation to individual user needs and creating interactive training environments.

One of the most promising applications of the GPT model is the simulation of therapeutic conversations. Its capacity to generate patient utterances with varying degrees of emotional complexity allows future therapists to practise interactions in realistic yet fully controlled settings. The model can reproduce different narrative styles and personality types, enabling the testing of diverse therapeutic strategies and the development of skills such as active listening, paraphrasing, or formulating appropriate questions. Unlike static textbooks, these simulations allow for flexible adjustment of difficulty levels, enabling users to progress from simple scenarios to more demanding clinical cases.

GPT's ability to generate text in real time makes it suitable for producing educational materials that reflect real clinical situations. This allows for the automatic creation of case studies, diagnostic tests, or detailed descriptions of therapeutic interventions. Psychology students can use these tools to analyse different therapeutic approaches, while experienced therapists may use them to test the effectiveness of new treatment methods. GPT can also assist academic instructors in designing interactive teaching materials tailored to the learners' knowledge levels and needs.

The model is also applicable in the assessment of therapeutic competence, allowing for linguistic analysis of psychology students' and therapists' utterances during simulated conversations. Through its advanced language processing mechanisms, GPT can identify key elements of effective communication, evaluate the correctness of techniques used, and suggest ways to improve patient interaction. The ability to generate personalised feedback reports enables users to consciously refine their

skills and adapt therapeutic strategies to the specific context of each conversation—especially in the development of empathic communication.

In the future, GPT-based models may become an integral component of psychologist and therapist training, serving as intelligent educational assistants. Owing to their ability to analyse language and generate realistic interactions, they can support both theoretical learning and practical preparation for clinical work. These technologies may also be used in therapeutic supervision, where the model would analyse session transcripts and provide feedback on the effectiveness of the interventions applied. Integrating GPT with psychological education may introduce a new quality into the learning process, making it more interactive, flexible, and responsive to the individual needs of each learner.

21.3.3. Psychological Aspects of Attachment to Artificial Intelligence

As artificial intelligence—particularly advanced language models—gains increasing importance in the field of psychology, the need arises to reflect on a phenomenon the author would describe not so much as *addiction*, but rather as *attachment to artificial intelligence*. This issue is being raised with growing frequency in both public and scientific discourse, and psychology, as a discipline that studies human–technology relations, will need to engage with it attentively.

The sources of such attachment may be rooted in at least two distinct types of experience: (a) interaction with another human being, and (b) interaction with a machine. In the case of human relations, we are dealing with a mechanism known in cybernetics as a feedback loop (Mazur, 1966). The result of such interaction is a dynamic sum of mutually reinforcing reactivities between object X and object Y (Mazur, 1976). To illustrate this with a simple example: person X responds to a stimulus and generates a message, which becomes a stimulus for person Y. However, this is not a "primary" stimulus—it has already been transformed by the reactivity of person X. Person Y then responds, processing that message through their own reactivity, and their response becomes a new stimulus that returns to person X—now multiplied by the reactivities of both X and Y. Person X then reacts again, with their response once more amplified by their own reactivity, creating a continuous loop of complex, often unpredictable, and even escalating feedback interactions.

The author explored this phenomenon in detail during a conference presentation and in one of her articles, analysing it in the context of regulating parent–child interaction under conditions of projective identification (Szymańska, 2024a). She emphasised that the stimulus perceived by the parent is not a "primary" one but has already been shaped by the child's reactivity. If we denote the original stimulus as $\mathbf{X}$ and the child's reactivity as r_a, the signal reaching the parent takes the form $\mathbf{Y_1} = \mathbf{X} \cdot r_a$. The parent processes it through their own reactivity r_b, resulting in the child receiving a stimulus $\mathbf{Y_2} = \mathbf{Y_1} \cdot r_b$. In response, the child reacts, processing this new signal again through r_a, which produces a further stimulus $\mathbf{Y_3} = \mathbf{Y_2} \cdot r_a$, and so on. In this way, a dynamic of mutually reinforcing reactivities is created—a classical

example of the feedback loop described in cybernetics (Mazur, 1966, 1976). This very instability, variability, and emotional ambiguity is often both the challenge and the depth of human relationships.

In interaction with AI, this mechanism is significantly disrupted—or, more precisely, suspended. According to the author, a person interacting with artificial intelligence experiences something fundamentally different: not so much contact with an "other" but a specific, pleasant form of contact with themselves. This arises from the fact that algorithms possess no inherent reactivity—they do not multiply stimuli through internal emotional, intentional, or contextual filters. They do not respond like a human being; rather, they replicate, amplify, and return—often with a high degree of predictability and coherence. The message that returns to the user does not contain resistance, is free of distortions caused by someone else's emotions, and does not trigger unpredictable reactions—it is almost "clean". One could say that a person conversing with AI receives back something that has been processed in an orderly, balanced manner, often explicitly focused on their cognitive or emotional needs. In this sense, attachment may not stem from a sense of the presence of another person but rather from the experience of a safe mirror that reflects the message without interference—and this may be psychologically highly appealing, especially for individuals who are hypersensitive or overwhelmed by interpersonal relationships.

The second quality that, according to the author, artificial intelligence offers to the human user is not only the effect of a "clean mirror". In conversations with AI, one does not experience dialogue in the classical interpersonal sense—instead, one engages in a kind of conversation with oneself. The absence of a feedback loop means that there is no second person introducing their reactivity, emotionality, or interpretation of stimuli. What AI "responds with" is not a response in the human sense—it is a mathematical transformation and composition of what the person has said, combined with the knowledge possessed by artificial intelligence. This knowledge, importantly, is not individual knowledge but a compressed repository of humanity's collective knowledge.

In this sense, when a person interacts with AI, they receive not only a clean response, but one that is free of another person's reactivity while simultaneously filtered through the vast body of knowledge embedded in the model. The AI's reply is, in simplified terms, a mathematical processing of the user's statement combined with information about the world. This means that in interacting with artificial intelligence, one stands, in a sense, before oneself—as if looking into a mirror that not only reflects one's words but amplifies them through access to an immense body of knowledge. Crucially, it does so without altering them through emotional filters. As a result, the person receives a message that is not coloured by "another human being". It is clean, calm, and nonjudgmental.

This is where the second layer of this experience lies. Artificial intelligence not only lacks reactivity—it also does not emotionally intervene in what a person says. If the user talks about a lawnmower, they receive information about a lawnmower. If they talk about the meaning of life, they receive a structured set of possible

answers to that question—but still devoid of a second person who reacts to their message with their own subjectivity. In this sense, AI does not conduct a conversation. It processes and returns. And ultimately, the person is engaging with themselves—and with knowledge. With knowledge that does not judge, provoke, or intrude. And that is precisely what makes these interactions so psychologically appealing: the possibility of inhabiting a cognitively safe space in which we receive something very rare—a message free from human reactivity, grounded in knowledge rather than emotion.

In summary, two main psychological mechanisms underlie—according to the author—the growing human attachment to artificial intelligence. The first is the absence of reactivity on the part of AI. In human interactions, feedback loops often generate tension, escalation, and difficult emotions. Every utterance is met with a response that carries not only content but also the emotional charge of the other person—their interpretation, resistance, reaction, and sometimes even rejection. In contact with AI, this component is absent. The response is neutral, clean, non-escalatory, and therefore non-burdensome. This creates the illusion of an ideal, peaceful dialogue space in which a person can speak—and be "heard"—without the risk of being hurt.

The second mechanism is the experience of being understood and the access to knowledge that requires no one's permission, approval, or consent. AI responds instantly, broadening the user's horizons and delivering information, structures, and meanings—unconditionally. For a person who carries a deep need to be understood, this can be an almost therapeutic experience. In a world where personal development often depends on another's willingness to engage, collaborate, or approve, AI becomes a tool for individual growth. It does not judge, block, or refuse. It simply responds. And it is precisely this combination—the absence of escalating reactivity and the experience of self-expansion through clearly structured, readily accessible knowledge—that forms the foundation of psychological attachment to artificial intelligence.

CHAPTER 22

Foundations of Text Mining: Tools and Techniques

22.1. Key Concepts and Methodologies in Text Mining

Text mining, as an interdisciplinary scientific field, employs techniques from natural language processing (NLP) and machine learning to analyse large volumes of textual data. Its primary aim is the extraction of relevant information and the transformation of unstructured data into actionable knowledge (Elder et al., 2012). This technology not only enables the identification of hidden patterns within data but also converts such data into forms that can be more deeply analysed and interpreted (Kulkarni & Mundhe, 2016; Szymańska, 2017b). This transformation is particularly valuable in the era of Big Data, where understanding and managing vast amounts of information is essential for progress across multiple scientific and industrial domains.

Text mining is not merely a technical process but also a methodology that dynamically shapes research approaches by offering new ways to understand the content and context of textual data. The integration of advanced algorithms, particularly those based on machine learning, facilitates the effective recognition, classification, and analysis of text. As a result, valuable information can be extracted from extensive datasets. These technologies enable the identification of trends, patterns, and relationships that are not visible to the naked eye, which is especially critical in contexts where rapid and precise data analysis can lead to significant scientific breakthroughs or competitive advantages (Kulkarni & Mundhe, 2016).

Text mining contributes to advancements in the analysis of textual data and its application across various scientific fields. It opens new perspectives for researchers in interpreting data. This interdisciplinary method combines both theoretical and practical approaches to text analysis, enabling a deeper understanding of structures and relationships hidden within content. The ability to identify patterns and trends that remain invisible without advanced textual analysis is one of the key advantages of text mining, making it an invaluable tool in modern analysis of unstructured data (Kulkarni & Mundhe, 2016).

22.1.1. Terms and Concepts Used in Text Mining

In text mining, fundamental terms such as tokenisation, stemming, lemmatisation, and sentiment analysis are essential for the effective processing of data. Each of these concepts plays a crucial role in transforming raw text into structures that can be further analysed, allowing for a deeper understanding of the context and meaning embedded in textual data.

Tokenisation refers to the process of dividing text into smaller units, such as words or sentences. It is the first and most fundamental step in textual analysis, enabling further operations to be performed on the text. By breaking the content into manageable segments, the text becomes more accessible for computational analysis. This process is necessary for more advanced algorithms to systematically process the data. Tokenisation is used in virtually every text mining project as a preliminary stage of data preparation (Elder et al., 2012).

Stemming is a technique for reducing words to their basic forms, or "roots". This allows different inflected forms of the same word (such as verb tenses or plural forms) to be treated as a single lexical item, facilitating both statistical and semantic analysis of the text. It helps standardise various word forms into a single representation. For example, words like "run", "running", and "ran" may be reduced to the root "run". Stemming is commonly used in information retrieval systems and recommendation engines, where it is important to group related content based on core word forms (Lovins, 1968).

Lemmatisation is similar to stemming but is a more advanced and precise process. It involves assigning words their canonical grammatical base forms, which allows for more accurate language processing (Manning & Schütze, 2002). Lemmatisation considers the context of a word, which is essential for understanding its proper meaning within a sentence. For instance, the Polish forms "leżał", "leży", and "leżą" would all be reduced to the lemma "leżeć". Lemmatisation is particularly useful in NLP tasks that require semantic precision, such as machine translation, educational applications of natural language processing, or dialogue systems where grammatical accuracy and contextual awareness are key.

Sentiment analysis is a technique used to assess the emotional tone of a text. It is particularly applied in social media analysis, where it can quickly and effectively determine whether user-generated content is positive, negative, or neutral. Sentiment

analysis allows for monitoring public opinion, consumer reactions to products, and general societal moods. For instance, companies may use this technique to track consumer responses to products or advertising campaigns by analysing comments and posts on social media platforms. In political analyses, it can be used to assess public sentiment toward candidates or parties.

These techniques form the foundation of all text mining processes, enabling the transformation of unstructured text into quantitative data that can be analysed in a more structured and measurable way (Elder et al., 2012; Szymańska, 2017b). They allow for deeper analysis of text and contribute to a better understanding of the information, trends, and patterns embedded within.

The application of these techniques facilitates the transformation of unstructured textual data into forms that can be systematically and quantitatively analysed, which is critical for the effective utilisation of information in both industrial and research contexts.

22.1.2. History and Development of Text Mining

The history of text mining is inseparably linked to the development of information technologies and artificial intelligence. The origins of text mining can be traced back to early information retrieval systems, which gradually evolved into advanced natural language processing (NLP) and machine learning technologies. The introduction of these technologies made it possible to process increasingly large volumes of textual data more efficiently, significantly expanding the capacity for analysis and interpretation. This, in turn, opened up new areas of application in business practice, scientific research, and data management (Hassani et al., 2020).

Text mining developed in parallel with the growing demand for methods capable of handling and analysing massive amounts of data generated in the digital world. From manual document indexing methods to automated data analysis systems, text mining has undergone substantial transformation over the past few decades. Initially focused on simple information extraction techniques, the field has gradually embraced more advanced methods of semantic analysis. Deep learning, neural networks, and other NLP techniques have become essential tools for researchers and analysts, enabling them to understand and utilise textual data in ways that were previously unattainable (Elder et al., 2012; Hassani et al., 2020).

The term "text mining" stems from the need to describe the process of "extracting" valuable knowledge from text. Although the exact year of its formal adoption is difficult to pinpoint, the term began to appear in academic literature as early as the 1980s and 1990s, in parallel with growing interest in information technologies. The term is often used interchangeably with "data mining", yet the distinctive feature of text mining lies in its exclusive focus on textual data for the purposes of analysis and processing (Elder et al., 2012).

The evolution of text mining also reflects broader developments in data collection and utilisation. As information technologies became increasingly sophisticated,

so too did the possibilities for applying text mining to process data on an unprecedented scale. This has contributed to the transformation of multiple sectors—from marketing and healthcare to finance—where text analysis offers new perspectives for understanding complex phenomena. Text mining has become not only an analytical tool but a critical component of information strategies, enabling deeper insight and more effective use of available textual data (Hassani et al., 2020).

CHAPTER 23

Introduction to Text Mining Applications in Psychology

Text mining, also referred to as text analysis, is the process of identifying patterns in unstructured data and transforming them into structured data that can subsequently be analysed using other algorithms, including artificial intelligence methods. The transformation of unstructured into structured data constitutes a key stage in the transition from text mining to data mining, where textual data, after preliminary processing, undergo deeper exploratory analysis (Elder et al., 2012). In psychology, where textual data are abundant, text mining can offer numerous benefits—from the analysis of interview and survey content to the examination of large datasets derived from social media. This chapter presents the significance of text mining in psychology, along with the AI tools and technologies used in the process.

23.1. The Importance of Text Mining in Psychology

Text mining in psychology aims not only at analysing textual data but also at uncovering hidden patterns and relationships that may be significant for understanding human behaviour and mental processes. Text analysis can be applied to the content of interviews—including clinical interviews—to reconstruct unstructured databases into structured ones and to analyse scientific literature.

Interview content analysis enables the identification of key topics and linguistic patterns that may be associated with particular mental states or disorders. By using text mining methods, it becomes possible to process large text corpora and extract information that might be overlooked in traditional qualitative analysis.

Text mining facilitates the automatic recognition and classification of topics and emotions expressed in statements—for example, those of patients. By analysing the content of clinical interviews, it is possible to identify frequently repeated words and phrases associated with anxiety, depression, or other psychological disorders. This allows therapists to better understand which topics and issues are most important to patients and what emotions accompany them (Ríssola, 2020).

*Transforming unstructured databases into structured ones***.** Text mining is a tool that enables the fast and efficient transformation of textual data into numerical data, which is particularly useful in psychology, where much of the data originates from interviews and narratives. Text mining algorithms enable word counting and weighting, analysis of word relationships using principal component analysis (PCA), and the conversion of verbal data into numerical form (Szymańska, 2017b). A series of studies has shown that text mining algorithms are useful for identifying hidden patterns and relationships between words in texts (Szymańska, 2019; Szymańska & Aranowska, 2016, 2019).

Scientific literature analysis using text mining techniques constitutes a valuable tool in psychological research. It enables the search of extensive publication collections, allowing for the rapid detection of research trends, underexplored areas, and new directions for investigation. As a result, psychologists can plan their studies more efficiently, focusing on the most promising issues and avoiding redundancy in previously conducted analyses.

Text mining supports the identification of popular topics within a given period by analysing keywords, titles, abstracts, and full texts of articles. Moreover, it enables the detection of knowledge gaps, allowing attention to be directed toward aspects that require further investigation. This tool also facilitates the discovery of innovative research topics that may make a significant contribution to the development of psychology.

The application of text mining optimises the literature search process, providing researchers with access to the most relevant and up-to-date publications. Recommendation algorithms can automatically suggest articles tailored to the specifics of ongoing research. As a result, these techniques support effective scientific information management and contribute to the dynamic advancement of psychology by identifying key research issues and directions for future exploration.

The advantages of using text mining include analysis speed and the ability to transform textual data into numerical data, which allows for subsequent statistical analyses, saving both time and research resources. The use of text mining in psychology can significantly reduce the time required for data analysis and enable the execution of advanced analyses that would be difficult to conduct using traditional methods (Szymańska, 2017b). Text mining algorithms may also be combined with other data mining methods, such as support vector machines, artificial neural networks, or *k*-means clustering, enabling an even more comprehensive approach to data analysis in psychology (Szymańska & Aranowska, 2019).

Text mining is widely used in psychological research, particularly in the analysis of various mental processes and linguistic patterns. An example of its application is the analysis of parental stress and personality traits shaped in children by parents experiencing varying levels of stress (Szymańska & Aranowska, 2019). The results of these studies confirmed the usefulness of text mining algorithms in identifying patterns and relationships in textual data (Szymańska & Aranowska, 2019).

Additionally, text mining has been applied to the analysis of narrative texts written by deaf children and to the analysis of individuals' internal dialogues, demonstrating the broad potential of this tool in psychology (Bokus et al., 2017; Jaworowska et al., 2016; Ważyńska et al., 2015). With text mining, it is possible to quickly identify patterns and dependencies in data, contributing to a better understanding of psychological and social processes.

The use of text mining algorithms in psychology can significantly enhance the analysis of textual data, aiding in the discovery of new patterns and relationships, thereby facilitating psychological research and increasing its efficiency (Szymańska, 2017b).

23.2. Turbo Text Mining: A Revolution in Text Exploration

With the emergence of modern language models such as BERT, GPT, and their derivatives, text analysis has reached a new level of efficiency and accuracy. Traditional text mining approaches were based primarily on statistical and linguistic methods which, while effective, were limited in their ability to account for context and understand complex linguistic structures. Next-generation language models have introduced mechanisms such as self-attention and transfer learning into text exploration, enabling the processing of massive volumes of text with exceptional precision. I have named this new stage of text analysis **Turbo Text Mining**, as it leverages the full spectrum of transformer model capabilities, allowing for deep, context-aware analysis of textual data. The term highlights a new quality in text exploration—speed, flexibility, and the ability to dynamically adapt to an evolving language.

Classical approaches to text exploration involved analysing word frequency, statistical topic modeling, and basic natural language processing techniques. A breakthrough occurred with the introduction of language models that not only recognise word sequences but also capture semantic dependencies between them in a broader context. This made it possible to automatically interpret user intent, extract hidden information from text, and perform more advanced sentiment and emotion analysis. In practice, this means that AI systems can now analyse content at the level of abstract concepts, rather than just individual words or phrases.

Modern text mining algorithms using language models are applied in a variety of fields—from opinion analysis in social media, through psychological diagnostics, to misinformation detection and legal content analysis. **Turbo Text Mining** enables dynamic topic modeling, meaning that algorithms are no longer confined to

rigidly predefined categories but can instead adapt to changing language and context. Thanks to precise semantic analysis, systems are capable of understanding the meaning of text in a way that resembles human reasoning, taking into account context and word ambiguity. This makes it possible to automatically generate summaries of long documents, where AI algorithms extract key information and present it in a concise, coherent format. Intelligent information retrieval powered by Turbo Text Mining enables systems to search knowledge bases more effectively and to identify answers to complex user queries based on contextual analysis.

One of the most important aspects of **Turbo Text Mining** is the ability of models to account for context and the dynamic nature of language. In traditional approaches, text analysis was often limited to static rules and dictionaries that required constant updating. Modern language models learn continuously and adaptively, allowing them to adjust to new linguistic trends, neologisms, or domain-specific contexts. **Turbo Text Mining** enables the exploration not only of content itself, but also of user interaction with the text, making it an invaluable tool in media monitoring, marketing analytics, and psychological research.

The use of language models in text analysis has led to a breakthrough in data exploration, creating a new standard in textual analysis. **Turbo Text Mining** is not only about faster and more accurate information extraction from text, but also about more advanced interpretation of content within social, psychological, and business contexts.

CHAPTER 24

Methods and Techniques Used in Text Analysis

Text mining employs both simple statistical methods and advanced machine learning algorithms to efficiently analyse texts. Basic statistical techniques, such as word frequency analysis, are useful for obtaining a general overview of the most frequently used terms within datasets. In contrast, advanced algorithms—such as neural networks and deep learning—enable more complex semantic analysis by identifying subtle patterns and dependencies in the data, which may be imperceptible to more traditional methods.

Talib and colleagues (2016) provide a detailed discussion of the various techniques used in text mining, including classification, clustering, and information extraction. Classification enables assigning texts to specific categories based on their content, which is crucial for information management and data filtering. Clustering allows for grouping similar texts without predefining categories, which is useful in exploratory data analysis. Information extraction refers to the process of retrieving specific data from texts, such as names, places, or dates, which is applicable in automatic summarisation or collecting contact data (Talib et al., 2016).

Classification is the process of assigning predefined labels to texts based on their content. It is one of the fundamental techniques used in information management and data filtering. For instance, in digital content management, classification can automatically categorise news articles into topics such as sports, politics, or culture. This technique is widely used in recommendation systems, where algorithms learn to recognise and classify content based on user preferences. Classification is key to the efficient search and organisation of large datasets (Gaikwad et al., 2014; Talib et al., 2016).

Clustering enables grouping texts without prior definition of categories, which is useful in exploratory data analysis. Clustering algorithms analyse texts and group them based on content similarity, which can help uncover hidden patterns or topics. This technique is invaluable, for example, in social media for identifying discussion trends or in academic literature for determining dominant research themes (Gaikwad et al., 2014; Talib et al., 2016).

Information extraction is the process of retrieving specific, structured information from unstructured textual data. Typical applications include the extraction of names, dates, locations, and other concrete elements, which can be used for automatic summarisation or for collecting contact data. This is particularly valuable in fields such as law and medicine, where rapid and accurate processing of large volumes of documents is of critical importance (Gaikwad et al., 2014).

Visualisation in text analysis is a dynamic and interactive process that significantly enhances the accessibility and comprehensibility of information extracted from large textual sources. This technique involves representing individual documents or document collections in visually attractive formats. Documents are often categorised using "text flags", which help to identify their category at a glance. Additionally, density colouring is used to indicate the volume or concentration of data in specific sections of the text, facilitating the detection of patterns and trends (Gaikwad et al., 2014).

Topic modeling enables the identification of latent topics within large text corpora, which can then be further analysed to understand dominant discourses. This technique is applied to trend analysis in social media or to the discovery of new research areas in scientific literature.

Named Entity Recognition (NER) is a natural language processing technique that identifies and classifies key elements in a text based on their category, such as person names, organisations, locations, temporal expressions, monetary amounts, percentages, and so forth (Liu et al., 2023). This technique is particularly useful for gathering and organising information from large text datasets, enabling the automatic extraction and indexing of relevant information. For example, in news articles, NER can be used to identify all mentioned persons, places, dates, and other specific data, thus facilitating the subsequent analysis of this information.

Semantic network analysis, in turn, is a method that goes a step further by analysing and visualising the relationships between the identified entities. A semantic network is a graph in which nodes represent entities (e.g., people, places, organisations) and edges describe the relationships between them (e.g., "works for", "lives in", "owns"). This type of analysis makes it possible to understand not only which entities appear in a text, but also how they are interconnected. For example, in the analysis of scientific texts, semantic network analysis can reveal how various concepts and theories are linked within a given field of knowledge, which may help identify key trends and research gaps.

These advanced techniques are extremely valuable in numerous practical applications—from automatic text summarisation and intelligent search systems to the

analysis of relationships and networks in social media and other forms of large textual data collections. By identifying and analysing relationships between entities, it becomes possible to gain a deeper understanding of the structure and meaning of texts, which is of crucial importance in the information age.

All these techniques make it possible to transform unstructured data into structures that can be more easily analysed and interpreted (Talib et al., 2016). As a result, a deeper understanding of texts is achieved, enabling the effective use of the information they contain for informed decision-making—ranging from market analysis and social media monitoring to supporting decision-making processes in various fields of the economy and science (Gaikwad et al., 2014).

24.1. Methodologies and Practical Applications of Sentiment Analysis

Sentiment analysis is one of the key applications of text mining, enabling the identification and classification of emotions contained within texts. It uses natural language processing (NLP) techniques to assess tone, opinion, and emotional content, with applications across various domains such as marketing, psychology, and social analysis. Within text mining, sentiment analysis is employed for the automatic extraction of user sentiment information, allowing for the detection of trends and the prediction of social responses.

Sentiment analysis techniques have evolved from traditional rule-based methods and machine learning classifiers to advanced deep learning models. In recent years, these methods have found widespread use in fields such as marketing, politics, psychology, and the monitoring of public opinion on social media platforms.

Traditional sentiment analysis methods include rule-based techniques, in which sets of linguistic rules are applied to identify sentiment, and machine learning approaches such as Support Vector Machines (SVM), which classify sentiment based on extracted text features (Latif et al., 2018a). Although these methods were effective in their time, their limitations become apparent in the face of the complexity of natural language and the need for scalable solutions.

Modern approaches to sentiment analysis include the use of neural networks such as Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Deep Belief Networks (DBN). These models are capable of automatically extracting relevant features and contexts from data (Latif et al., 2018b; Retta et al., 2023).

Sentiment analysis has wide-ranging applications in practice. In marketing, it is used to monitor consumer opinions and analyse product reviews, allowing for more effective adaptation of marketing strategies. In politics, sentiment analysis enables the monitoring of public opinion on politicians and policies, which is essential for conducting successful political campaigns. In psychology, the application of sentiment analysis in studies of mental health and emotional states opens up new diagnostic and therapeutic possibilities.

24.1.1. Techniques Used in Sentiment Analysis

Feature extraction techniques play a crucial role in sentiment analysis. Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech emotion analysis, representing short-term speech spectra that are relevant for identifying emotions (Retta et al., 2023). The acoustic feature set eGeMAPS is also commonly applied in speech emotion analysis (Retta et al., 2023). In text analysis, a significant development has been the creation of methods that allow for effective mapping of words into vector space (Mikolov et al., 2013).

Long Short-Term Memory (LSTM) is a type of neural network capable of processing sequential data, such as text, by capturing long-term dependencies between words (Retta et al., 2023). Convolutional Neural Networks (CNN) are used to automatically extract features from text (Retta et al., 2023). Transformer-based models, such as BERT and XLM, employ attention mechanisms to understand contexts and relationships in text, enabling more precise sentiment analysis.

24.1.2. Emotion Analysis in Social Media

Social media has become a primary data source for emotion analysis, enabling real-time monitoring of public mood and opinion. Emotion analysis in social media requires advanced natural language processing techniques and scalable tools for analysing large datasets.

Transformer-based models such as BERT, RoBERTa, mBERT, and XLM-R are widely used in social media sentiment analysis, allowing for effective processing of multilingual data. Real-time sentiment analysis is made possible by leveraging data streaming and Big Data technologies, which enable continuous monitoring and analysis of sentiment.

Brand monitoring on social media involves analysing opinions and sentiments associated with brands on platforms such as Twitter, Facebook, and Instagram, allowing companies to respond quickly to changing consumer attitudes. In the context of political campaigns, emotion analysis allows for the monitoring of public responses to campaigns and political events, which is critical for implementing effective electoral strategies. In scientific research, emotion analysis is used to study social and psychological behaviours, opening new possibilities in the fields of sociology and psychology.

Dealing with irony and sarcasm remains a challenge for NLP models, which must detect subtle forms of communication. Multilinguality presents another challenge, requiring effective sentiment analysis in the context of diverse languages and cultures. Protecting user data privacy is a key concern in the analysis of social media data, necessitating the implementation of appropriate safeguards.

CHAPTER 25

Topic Modeling Strategies and Semantic Network Analysis

Topic modeling and semantic networks are two distinct approaches to text analysis. While topic modeling enables the detection of dominant themes within a collection of documents, semantic networks allow for the analysis of relationships between concepts and the semantic structure of the text.

Topic modeling is a technique used in natural language processing (NLP) to identify latent topics within text corpora. It is particularly useful in the analysis of large textual datasets, such as scientific articles, social media posts, interview transcripts, and more. In psychology, topic modeling can assist researchers in understanding which themes dominate discussions on a given subject, which issues are most frequently addressed, and how these evolve over time. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA), enable researchers to uncover the thematic structure of documents without prior knowledge of their content.

Semantic networks are an important tool in psychological research, allowing for the analysis of the structure and dynamics of concepts and their interrelations. With advanced techniques in natural language processing (NLP) and data analysis, researchers can understand how individuals organise knowledge, what their associations are, and how these structures influence behaviour and thinking. Semantic networks are graphs in which nodes represent concepts and edges between them indicate semantic relationships such as similarity, synonymy, or antonymy.

25.1.1. Topic Modeling Techniques – LDA

One of the most commonly used methods in topic modeling is Latent Dirichlet Allocation (LDA). LDA is a generative statistical model that assumes documents

are mixtures of topics, and topics are mixtures of words. The process begins with an initialization step, in which the LDA algorithm assumes a fixed number of topics K, and each document is represented as a mixture of these topics. Initially, words are randomly assigned to topics, enabling the initial estimation of distributions (Blei et al., 2003). The algorithm then iteratively assigns topics to words in documents, maximising the likelihood of assigning words to topics. In this step, the Dirichlet distribution is used to model the mixture of topics. The model assumes that each document is generated by a mixture of topics, and each topic is a probability distribution over words.

After assigning topics to words, the algorithm updates the topic distribution for each document and the word distribution for each topic to better fit the model to the data. This update is performed by computing the conditional probabilities of assigning words to topics and topics to documents. The process is repeated until the model reaches a stable state in which the topic assignments of words do not significantly change across iterations. Stopping criteria typically include a maximum number of iterations or a minimal change in log-likelihood (Blei et al., 2003).

25.1.2. Topic Modeling in Text Analysis

Topic modeling is extremely useful in textual analysis within psychology. Researchers can apply topic modeling to the analysis of therapy session transcripts in order to identify the main themes raised by patients. This can help in understanding which issues are most important to patients and how these evolve over time. For example, thematic analysis of scientific articles over the years may reveal how research priorities have shifted and which new topics have gained prominence (Blei & Lafferty, 2007).

Topic modeling can also be used to analyse narratives in studies of memory and identity. Researchers can examine which themes appear in individuals' stories about their past and how these themes are related to their identity and experiences. In social psychology, topic modeling can be applied to the analysis of social media posts to understand which topics are most discussed within different social groups and how these discussions influence social behaviours. By applying LDA, researchers can study topic shifts over time, allowing for a better understanding of social and psychological dynamics.

Topic modeling, and particularly LDA, is a powerful tool in textual analysis that enables the automatic discovery of thematic structures within large textual datasets. In psychology, these techniques can significantly enhance our understanding of human thought, behaviour, and social interaction.

Beyond topic modeling, another important method of text analysis in psychology is semantic network analysis, which allows for the investigation of relationships between concepts and the semantic structure of language.

25.2. Construction and Analysis of Semantic Networks in Psychological Research

Semantic networks allow for the structural analysis of relationships between concepts, which can lead to a deeper understanding of cognitive and emotional processes. These networks are particularly useful for modeling knowledge representation in the human mind, studying semantic memory, and conducting linguistic analysis. For example, the analysis of word associations and their organisation within a network can provide insight into how people process information and how different concepts are interconnected in the mind (Collins & Quillian, 1969).

Semantic networks can also be applied to the study of complex cognitive processes such as reasoning, problem-solving, and decision-making. By analysing patterns of relationships between concepts, researchers can identify key elements and structures that support these processes (Collins & Quillian, 1969). For instance, understanding how people categorise information and construct conceptual hierarchies can contribute to the development of more effective teaching methods and therapeutic interventions.

In addition, semantic networks are used in the analysis of language and communication, enabling researchers to study how different social and cultural groups create and interpret meaning. Semantic network analysis can help identify linguistic patterns characteristic of specific groups, which is relevant in sociolinguistic and cross-cultural studies (Puertas et al., 2021).

Thanks to NLP techniques such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014), ontologies, and dependency analysis, it is now possible to automatically build and analyse semantic networks on a large scale. These advanced tools allow for more efficient processing of vast textual datasets, opening new research opportunities in psychology and other social sciences.

25.2.1. Methods for Building Semantic Networks

Traditionally, semantic networks were constructed manually by experts who defined concepts and the relationships between them. While this method ensures high quality and precision, it is time-consuming and subjective. The automatic construction of semantic networks has gained importance with the development of NLP and machine learning techniques, which allow for building networks based on large textual datasets.

One of the key techniques is information extraction, which includes processes such as tokenisation, part-of-speech tagging, and dependency parsing. Tokenisation splits the text into individual words or phrases; part-of-speech tagging assigns grammatical labels to each word; and dependency parsing identifies grammatical relationships between words.

Word embeddings, such as word2vec, GloVe, and BERT, represent words as vectors in a multidimensional space. This allows for the analysis of semantic similarity

between concepts, enabling the automatic discovery of semantic relationships in large text datasets (Mikolov et al., 2013; Pennington et al., 2014).

Cluster analysis, using algorithms such as *k*-means and hierarchical clustering, groups concepts based on their similarity, helping to identify structures within the semantic network.

25.2.2. Applications of Semantic Networks in Psychological Research

Semantic networks are used to study how people organise their knowledge and what cognitive structures influence their thinking. These studies may include the analysis of word associations, semantic memory, and categorisation processes. For example, semantic networks can help to understand how different concepts are connected in the mind (Collins & Quillian, 1969), as well as how these connections affect decision-making and problem-solving processes.

Semantic networks can also be used in clinical psychology to analyse patients' thought patterns and identify cognitive distortions. The analysis of these networks may assist in diagnosing and treating disorders such as depression or anxiety. For instance, by analysing a patient's thinking patterns, therapists can better understand the sources of negative thoughts and propose effective therapeutic interventions.

Semantic network analysis is also applied in studies of language and communication, enabling the investigation of how people create and interpret meaning in a social context. Examples include the analysis of narratives, discourse, and intercultural communication. Semantic networks can help identify linguistic patterns that are characteristic of different social or cultural groups, which may be relevant in sociolinguistic research.

CHAPTER 26

Ethical Aspects and Privacy in the Text Mining Process

26.1. Ethical Challenges and Privacy

The introduction of advanced language models, such as GPT-3, presents a number of ethical challenges. GPT-3, as one of the most powerful NLP models, can generate texts that are virtually indistinguishable from those written by humans (Dehouche, 2021). This capability gives rise to serious concerns, including the potential generation of harmful content, misinformation, spam, phishing attempts, and misuse in legal and governmental processes.

Phishing is a type of fraud in which the attacker impersonates a trusted person or institution to obtain confidential information such as passwords, credit card numbers, or other personal data. Scammers often use fake emails, text messages, or websites that appear legitimate to trick victims into revealing their data. Language models such as GPT-3 can be used to generate more convincing phishing messages, thereby increasing the risk of such attacks.

Moreover, there is a risk that these models could be used to produce fake scientific articles or academic essays, which poses a threat to the integrity of scientific research and education.

However, it is also important to recognise the benefits of using such models. Working with GPT or other advanced language models is not a one-sided process. Researchers who use these tools actively contribute to the process of data analysis and interpretation. Collaboration between humans and algorithms can lead to the discovery of new perspectives and the generation of innovative ideas that would be

difficult to achieve otherwise. This synergistic approach—combining human creativity and intuition with the data-processing capabilities of algorithms—can significantly advance many areas of science and technology.

In the process of text mining, data often comes from publicly available sources such as social media, internet forums, or blogs. Despite the public nature of such data, the privacy of individuals to whom the data pertains remains critical. **Anonymisation** is one method used to protect privacy; however, it is not always sufficient. There is a risk of re-identifying individuals based on anonymised data, especially when it is combined with other datasets. Therefore, anonymisation and pseudonymisation techniques must be applied with the utmost care to ensure full privacy protection.

Language models like GPT can reflect and amplify existing biases present in their training data. This can lead to discrimination against certain social groups, posing a serious ethical challenge. For example, recruitment systems may unintentionally favor candidates of a particular gender or ethnic background (Bender et al., 2021). It is essential to regularly monitor and test models for bias and apply techniques that minimize such biases to ensure fair and reliable outcomes.

One of the greatest ethical challenges associated with GPT is the issue of **plagiarism**. GPT can generate texts that are difficult to distinguish from original works written by humans, raising questions about intellectual property and authorship (Dehouche, 2021). Who is the author of a text generated by AI? Is it the person who posed the prompt, the team responsible for developing the model, or the model itself? These questions are particularly important in academic contexts, where the attribution of authorship is essential.

On the other hand, collaboration with such models can be highly beneficial. Users of these tools actively participate in the creative process by providing input that is then processed by the algorithm. This interactive approach enables the exploration of new ideas and concepts, which can subsequently be developed and refined by the human user. In this way, GPT and similar models can be viewed as assistive tools that expand the capacities of human thought and creativity.

26.2. Principles of Responsible Data Use

Transparency and **reproducibility** are fundamental to the credibility of research. In the context of text mining, this includes sharing source code, datasets, and analysis results. As Mittelstadt et al. (2016) note, a lack of transparency in algorithmic processes can lead to serious ethical consequences—observations that also apply to text mining practices. Therefore, researchers should strive for maximum transparency to enable verification and replication of their studies.

Education and **training** of researchers and practitioners in the areas of ethics and privacy are essential. The authors highlight the need to invest in skill development and workforce retraining in the face of automation (Wright & Schultz, 2018),

which can be extended to include ethical education for individuals designing and implementing AI systems. Introducing best practices and educational resources can significantly raise awareness in areas such as model bias, data anonymisation, and user consent.

The FAIR principles—**Findability**, **Accessibility**, **Interoperability**, and **Reusability**—are essential for responsible data management. Adhering to these principles in the context of text mining can promote greater transparency and usability of data, allowing it to be reused in future research. Striving for FAIR compliance should be standard practice in all research activity.

Researchers must be aware of their responsibility for the outcomes of text mining analyses. It is essential that the results are **reliable** and **fair**, and that researchers remain aware of **the ethical implications** of their work (Mittelstadt et al., 2016). Avoiding and minimising **bias in models** is crucial to ensure that the outcomes are not discriminatory.

26.3. Ethical Dilemmas of Artificial Intelligence: Is the Non-Use of AI Morally Justifiable?

In the age of artificial intelligence (AI), the ethical question is no longer limited to how AI will change the law or how to use it responsibly, but also whether it is ethical *not* to use it. Artificial intelligence holds enormous potential to advance science, support children's learning, and level the playing field for gifted children from underprivileged backgrounds who lack educational support.

AI can significantly accelerate research processes by automating data analysis and identifying new research directions. For example, AI can assist in processing massive datasets, thereby speeding up scientific discoveries and increasing the efficiency of research.

In education, AI has the potential to personalise learning by adapting materials to the needs and capabilities of each student (Holmes et al., 2019). AI-based systems can monitor students' progress and provide personalised recommendations, which can be especially helpful for gifted children from low-income families. AI could be the key to creating a more equitable world, in which every child, regardless of their background, has an equal opportunity to succeed. Ignoring this potential may be as unethical as the misuse of the technology itself. AI has the capacity to reduce educational inequality, especially for children from disadvantaged backgrounds who lack access to high-quality educational materials or academic support.

Undoubtedly, AI has the potential to contribute to the development of social goods, as emphasised by contemporary researchers (Shi et al., 2020). This raises an important question: if we are aware of how greatly AI can enhance science and education, is it ethical not to use it? By choosing not to use AI, are we depriving children of opportunities that could shape their future? An ethical approach to AI should

not be limited merely to legal regulation and responsible use but should also consider the societal impact of *not* employing such technology.

Ethical deliberations concerning artificial intelligence should encompass both regulatory dimensions and the moral obligations associated with its use. AI has the potential to yield profound social, scientific, and educational benefits, and failing to utilise this technology may be just as problematic as its unethical application. Therefore, the question of whether it is ethical to refrain from using AI is equally important and warrants serious reflection.

CHAPTER 27

Workshops and Practical Applications of Text Mining

Text mining—techniques for processing and analysing textual data—is gaining increasing popularity in psychological research due to its ability to transform large text corpora into meaningful quantitative information. The integration of these techniques into everyday research practice enables psychologists to identify hidden patterns, analyse the content of clinical interviews, study social media discourse, and explore scientific literature. This chapter discusses practical applications of text mining aimed at enhancing researchers' skills in analysing textual data and applying these techniques in the field of psychology. Although modern language models such as BERT and GPT offer immense potential, this chapter deliberately focuses on classical techniques that form the essential foundation for understanding more advanced NLP tools.

The first topic addressed is word frequency analysis. This is one of the most fundamental text mining techniques, allowing researchers to identify the most common words and phrases in a given corpus. This method provides an initial understanding of the structure of the text and helps identify key themes relevant to further analysis. Workshops on word frequency analysis cover word counting techniques, dictionary creation, frequency chart generation, and interpretation of results in psychological research contexts.

Another essential topic is the transformation of qualitative data into quantitative data. This is a key component of text mining that allows for more advanced statistical analyses. Methods such as thematic coding**,** sentiment analysis, and principal component analysis (PCA) enable the conversion of unstructured textual data into numerical data that can be further analysed using statistical tools. Workshops in this

area include an overview of transformation methods, practical application examples, and discussion of software tools that support this process.

The final, but no less important, topic is the integration of qualitative and quantitative analyses in psychological research. Integrating these two approaches allows for a more complete understanding of the studied issues by combining detailed descriptions and interpretations (qualitative analysis) with broad patterns and statistical generalisations (quantitative analysis). This combination enables researchers to obtain a more comprehensive understanding of psychological phenomena. Workshops on integrating qualitative and quantitative analyses include a review of mixed-methods methodologies, techniques for combining data from various sources, and practical examples from psychological research.

All the topics presented in this chapter aim to equip researchers with the essential skills and knowledge necessary to effectively use text analysis techniques in both their scientific and applied work. In the following sections, we will explore these techniques in detail, presenting both their theoretical foundations and practical applications.

Modern language models such as BERT and GPT offer immense capabilities in text processing; however, their application is not always optimal in scientific and didactic contexts. In this chapter, I deliberately focus on classical text mining methods, as I aim to introduce readers to the foundational techniques for transforming unstructured data into structured formats. This is a crucial step for researchers who intend to apply various data analysis algorithms, including classification, prediction, clustering, and regression methods.

The goal of this chapter is to teach readers how to think about texts as datasets and to guide them in the process of moving from raw text to its structured representation. Only with this understanding can more advanced techniques be used consciously and effectively. This book discusses various classes of algorithms—clustering, association analysis, decision trees, predictive models (e.g., SVM), *k*-means algorithms, and many others. To apply these methods effectively, one must work with data that is appropriately structured.

In the classical approach I present, the researcher maintains full control over the process: they can select feature extraction methods, adjust analysis parameters, and experiment with different techniques. Such an approach is more valuable for those who wish to analyse textual data consciously rather than simply rely on ready-made tools. The aim of this chapter is to provide readers with the skills to independently transform unstructured data into structured data and to integrate that data with other analytical methods.

After completing this process, the reader will be equipped to make informed use of more advanced NLP models such as BERT, with full control over the data structure and an understanding of the transformations occurring at each stage of analysis. The goal is not to simplify text analysis, but to equip researchers with a deep understanding of the process—so that they are not merely users of off-the-shelf models, but are capable of consciously building structured datasets and integrating them with other analytical techniques.

27.1. Word Frequency Analysis in Text

This chapter focuses on the practical applications of text mining in the analysis of textual data in psychology. It discusses how algorithms can be used to count words in textual documents and how Principal Component Analysis (PCA) can be applied for preliminary exploration of such data. Word frequency analysis is one of the fundamental techniques in text mining. It enables the identification of the most frequently occurring words and phrases within a corpus, allowing for an initial understanding of the structure of the text and the identification of key themes. Later sections of the chapter demonstrate how PCA techniques can be employed to identify hidden patterns and relationships among words, thus facilitating more advanced statistical analyses. The chapter also presents supporting software tools and methods to enable researchers to apply these techniques effectively in their own studies.

As an example, the chapter draws on an analysis of parental goals adopted by mothers toward their children. The aim of the study was to examine what parental goals mothers seek to cultivate in their children, and whether similar goals are chosen for sons and daughters. *Parental goals* refer to the psychological traits that parents wish to develop in their children, and they can be divided into desirable traits that parents aim to promote and undesirable traits that they attempt to suppress.

Similar analyses have been published previously, but since the current aim is to illustrate the potential of text mining, the analysis was intentionally made more complex (Szymańska & Aranowska, 2023). In addition to general parental goals, the analysis was conducted separately for goals mentioned in various positions on the list, rather than aggregating all mentions as in prior studies. Traits listed in the first, second, and third positions were considered separately, as well as the three traits marked as undesirable—those that parents strive to prevent from developing in their children.

The analyses were performed in two separate groups: mothers of boys and mothers of girls. This allowed for the investigation of how mothers articulate their parental goals depending on the child's gender. By counting the frequency of each trait and applying PCA to identify patterns, it was possible to gain a better understanding of the values and attitudes mothers aim to pass on to their children.

In the group of mothers of girls, the algorithms identified nine traits most commonly listed in the first position (see Table 27.1 and Figure 27.1, panel a). The most frequently mentioned trait in the first position was *independence* (mentioned 45 times), followed by *kindness* (9 times), *self-confidence* (8 times), *courage* (7 times), *empathy* (6 times), and five traits mentioned five times each: *autonomy*, *happiness*, *values*, and *faith*.

Table 27.1. Psychological Trait Matrix of Mothers in Relation to Girls

ID	Goal	Order	Count	Gender
independence	desirable	first	45	girl
kindness	desirable	first	9	girl
self-confidence	desirable	first	8	girl
courage	desirable	first	7	girl
empathy	desirable	first	6	girl
autonomy	desirable	first	6	girl
happiness	desirable	first	5	girl
values	desirable	first	5	girl
faith	desirable	first	5	girl
independence*	desirable	second	14	girl
empathy*	desirable	second	11	girl
creativity*	desirable	second	8	girl
faith*	desirable	second	7	girl
assertiveness*	desirable	second	6	girl
openness*	desirable	second	5	girl
courage*	desirable	second	5	girl
respect*	desirable	second	5	girl
world*	desirable	second	5	girl
honesty*	desirable	second	5	girl
skill*	desirable	second	5	girl
sensitivity*	desirable	second	5	girl
empathetic**	desirable	third	17	girl
courage**	desirable	third	8	girl
creativity**	desirable	third	6	girl
confidence**	desirable	third	6	girl
independence**	desirable	third	6	girl
assertiveness**	desirable	third	5	girl
intelligence**	desirable	third	5	girl
honesty**	desirable	third	5	girl
selfishness§	undesirable	fourth	17	girl
laziness§	undesirable	fourth	12	girl
aggressiveness§	undesirable	fourth	11	girl
malice§	undesirable	fourth	7	girl
pessimism§	undesirable	fourth	5	girl
selfishness	undesirable	fifth	15	girl
aggression	undesirable	fifth	8	girl
laziness	undesirable	fifth	6	girl
submissiveness	undesirable	fifth	6	girl
world§	undesirable	fifth	6	girl
laziness	undesirable	sixth	10	girl
selfishness	undesirable	sixth	7	girl

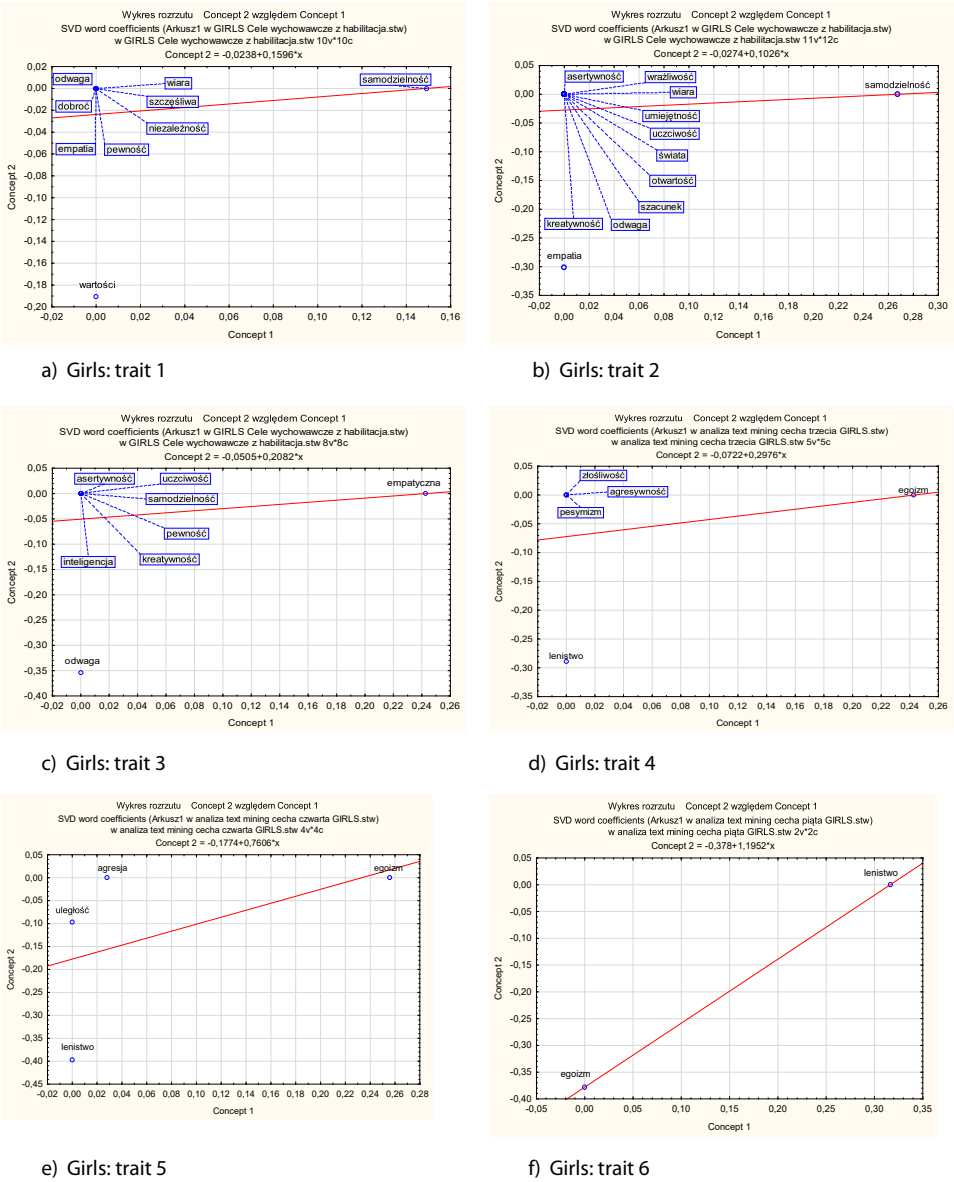


Figure 27.1. Results of Principal Component Analysis for Parental Goals Adopted by Mothers Toward Daughters

Principal component analysis revealed that *independence* and *values* formed two separate clusters, i.e., they did not co-occur either with each other or with any of the other traits. The remaining traits formed a shared cluster, meaning they were frequently mentioned together by mothers (Figure 27.1, panel a). This raises the

question: with which traits did *independence* and *values* actually co-occur? In reality, mothers listed a large variety of traits; however, for the purposes of this analysis, only those mentioned at least five times are included.

In the second position, the most frequently mentioned trait was again *independence* (14 mentions), followed by *empathy* (11 mentions), *creativity* (8), *faith* (7), *assertiveness* (6), *openness* (6), and five mentions each of *courage*, *respect*, *curiosity about the world*, *honesty*, *capability*, and *sensitivity*. Principal component analysis indicated that *independence* and *empathy* formed two separate clusters, i.e., they did not co-occur with each other or with other traits. The remaining traits formed a shared cluster (Figure 27.1, panel b).

In the third position, mothers of daughters most frequently listed *empathy* (17 mentions), followed by *courage* (8), *creativity* (6), *confidence* (6), *independence* (6), and five mentions each of *assertiveness*, *intelligence*, and *honesty*. Principal component analysis revealed that *empathy* and *courage* formed two separate clusters, i.e., they did not co-occur with each other or with other traits. The remaining traits formed a shared cluster (Figure 27.1, panel c).

The fourth, fifth, and sixth sets of traits reflect the undesirable characteristics that mothers aim to suppress in their daughters. Among the fourth-position traits, the most undesirable was *selfishness* (17 mentions), followed by *laziness* (12), *aggressiveness* (11), *malice* (7), and *pessimism* (5). Principal component analysis indicated that *selfishness* and *laziness* formed two distinct clusters, i.e., they did not co-occur with each other or with other traits. The remaining traits were grouped into a shared cluster (Figure 27.1, panel d).

Among the fifth-position traits, the most undesirable was again *selfishness* (15 mentions), followed by *aggression* (6), *laziness* (6), and *submissiveness* (6). Principal component analysis revealed that all of these traits formed separate clusters, meaning they did not co-occur with each other (Figure 27.1, panel e).

Among the sixth-position traits, the most undesirable was *laziness* (10 mentions), followed by *selfishness* (7 mentions). Principal component analysis indicated that these traits formed two distinct clusters, i.e., they did not co-occur with each other (Figure 27.1, panel f).

In the group of mothers of boys, the algorithms identified ten traits most frequently mentioned in the first position (see Table 27.2 and Figure 27.2, panel a). The most commonly mentioned trait was *independence* (41 mentions), followed by *empathy* (10), *self-confidence* (8), *creativity* (7), *courage* (6), *obedience* (6), *faith* (6), *assertiveness* (5), *intelligence* (5), and *honesty* (5). Principal Component Analysis (PCA) for the group of mothers of boys revealed that *independence* and *empathy* formed two separate clusters, meaning they did not co-occur either with each other or with any of the other traits. The remaining traits formed a shared cluster, indicating that they were often mentioned together by the mothers (Figure 27.2, panel a).

In particular, traits such as *courage*, *intelligence*, *respect*, *honesty*, *faith*, and *creativity* tended to co-occur, suggesting their association in the context of the developmental goals mothers set for their sons. These findings show that mothers have

a tendency to group certain traits together, recognizing them as complementary in shaping boys' personalities.

Independence and *empathy*, on the other hand, were listed as distinct, standalone educational goals. This suggests that mothers view them as important yet independent characteristics, not necessarily connected with other traits. In reality, mothers listed a wide range of traits not included in this analysis, which highlights the complexity and diversity of their parenting goals.

These results indicate significant differences in how mothers approach the shaping of their sons' personalities, emphasizing unique priorities and values attributed to various traits.

In the second position, the most frequently mentioned trait was again *independence* (14 mentions), followed by *empathy* (11), *resourcefulness* (8), *intelligence* (6), *creativity* (6), *sensitivity* (6), *assertiveness* (5), *kindness* (5), *courage* (5), and *respect* (5). PCA of the traits listed in the second position by mothers of boys showed that *independence* and *empathy* again formed two separate clusters, meaning they did not co-occur with the other traits. The remaining traits—including *resourcefulness*, *intelligence*, *creativity*, *sensitivity*, *assertiveness*, *kindness*, *courage*, and *respect*—formed a shared cluster, indicating that they were often mentioned together by mothers (Figure 27.2, panel b).

In particular, traits such as *resourcefulness*, *intelligence*, *creativity*, *sensitivity*, *assertiveness*, *kindness*, *courage*, and *respect* tended to co-occur, indicating their interrelation in the context of parenting goals set by mothers for their sons. These findings suggest that mothers often group certain traits together, perceiving them as complementary in fostering the development of boys' personalities.

By contrast, *independence* and *empathy* were listed as distinct, standalone parenting goals, which suggests that mothers view these traits as important and independent, but not necessarily associated with other characteristics. In reality, mothers listed a wide variety of traits not included in this analysis, highlighting the complexity and diversity of their parenting objectives.

In the third position, the most frequently mentioned trait was *empathy* (16 mentions), followed by *honesty* (7), *self-confidence* (6), *independence* (6), *creativity* (5), *responsibility* (5), *courage* (5), *optimism* (5), and *openness* (5). Principal Component Analysis (PCA) for the traits listed in the third position by mothers of boys revealed that *empathy* and *honesty* formed separate clusters, meaning they did not co-occur with other traits. The remaining traits—including *courage*, *self-confidence*, *creativity*, *openness*, *optimism*, and *responsibility*—formed a shared cluster, suggesting that they were often mentioned together by mothers (Figure 27.2, panel c).

In particular, traits such as *courage*, *self-confidence*, *creativity*, *openness*, *optimism*, and *responsibility* tended to co-occur, pointing to their association in the context of developmental goals set by mothers for their sons. These results again show that mothers are inclined to group certain traits together, perceiving them as synergistic in shaping boys' character.

In contrast, *empathy* and *honesty* were again listed as separate, distinct goals, indicating that mothers regard them as individually important but not necessarily related to other traits. As in previous cases, a much broader range of traits was mentioned by mothers but excluded from this analysis, reflecting the complexity and richness of parenting strategies.

The fourth, fifth, and sixth traits are undesirable characteristics whose development is suppressed by mothers. Among the fourth traits, the most undesirable one in the group of mothers of boys is *egoism*, mentioned 18 times, followed by *aggressiveness* (18 times), *laziness* (13 times), and *malice* (7 times). Principal component analysis (PCA) for undesirable traits listed first by mothers of boys revealed that *aggressiveness* and *egoism* formed separate clusters, which means they did not co-occur with other traits. The remaining traits, such as *laziness* and *malice*, formed a joint cluster, suggesting that they were often mentioned together by mothers (Figure 27.2, panel d).

In particular, traits such as *laziness* and *malice* co-occurred, indicating their association in the context of traits that mothers do not wish to develop in their sons. These results show that mothers tend to group certain undesirable traits together, perceiving them as particularly detrimental to the development of boys' personalities.

Aggressiveness and *egoism*, in contrast, were listed as separate, distinct undesirable goals, which suggests that mothers consider them exceptionally harmful and important to avoid, regardless of other traits.

Among the fifth traits, the most undesirable one is *aggression*, mentioned 14 times, followed by *laziness* (11 times) and *egoism* (8 times). Principal component analysis (PCA) for undesirable traits listed second by mothers of boys revealed that *aggression*, *egoism*, and *laziness* formed separate clusters, meaning they did not co-occur with other traits (Figure 27.2, window e). All three traits were listed as distinct undesirable goals, suggesting that mothers perceive them as particularly harmful and important to avoid.

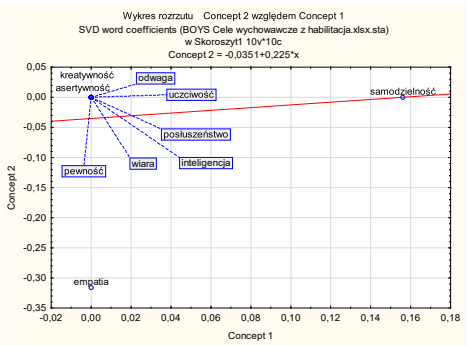
In particular, *laziness* was the most frequently mentioned trait, indicating its high relevance in the context of characteristics that mothers do not wish to foster in their sons.

Among the sixth traits, the most undesirable one is *laziness*, mentioned 9 times, followed by *egoism* (6 times) and *pessimism* (5 times). Principal component analysis (PCA) for undesirable traits listed third by mothers of boys revealed that *laziness*, *egoism*, and *pessimism* formed separate clusters, meaning they did not co-occur with other traits (Figure 27.2, window f). All three traits were listed as distinct undesirable goals, suggesting that mothers consider them particularly harmful and important to avoid.

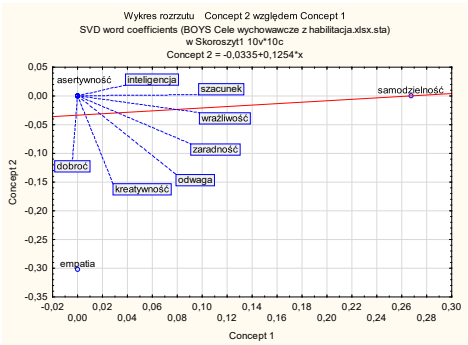
In particular, *egoism* was the most frequently mentioned trait, indicating its high relevance in the context of characteristics that mothers do not wish to foster in their sons.

Table 27.2. Psychological Trait Matrix of Parents in Relation to Boys

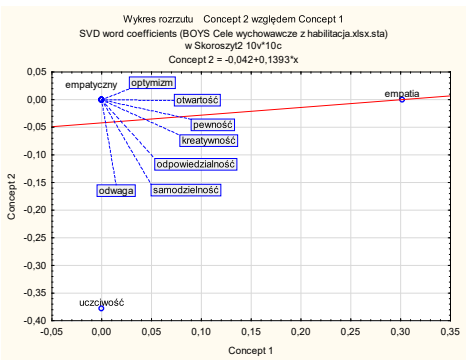
ID	Goal	Order	Count	Gender
independence	desirable	first	41	boy
empathy	desirable	first	10	boy
self-confidence	desirable	first	8	boy
creativity	desirable	first	7	boy
courage	desirable	first	6	boy
obedience	desirable	first	6	boy
faith	desirable	first	6	boy
assertiveness	desirable	first	5	boy
intelligence	desirable	first	5	boy
honesty	desirable	first	5	boy
independence*	desirable	second	14	boy
empathy*	desirable	second	11	boy
resourcefulness*	desirable	second	8	boy
intelligence*	desirable	second	6	boy
creativity*	desirable	second	6	boy
sensitivity*	desirable	second	6	boy
assertiveness*	desirable	second	5	boy
kindness*	desirable	second	5	boy
courage*	desirable	second	5	boy
respect*	desirable	second	5	boy
empathy**	desirable	third	16	boy
honesty**	desirable	third	7	boy
confidence**	desirable	third	6	boy
independence**	desirable	third	6	boy
creativity**	desirable	third	5	boy
responsibility**	desirable	third	5	boy
courage**	desirable	third	5	boy
optimism**	desirable	third	5	boy
openness**	desirable	third	5	boy
selfishness§	undesirable	fourth	18	boy
aggressiveness§	undesirable	fourth	18	boy
laziness§	undesirable	fourth	13	boy
malice§	undesirable	fourth	7	boy
laziness	undesirable	fifth	11	boy
selfishness	undesirable	fifth	8	boy
aggression	undesirable	fifth	14	boy
selfishness	undesirable	sixth	6	boy
laziness	undesirable	sixth	9	boy
pessimism	undesirable	sixth	5	boy



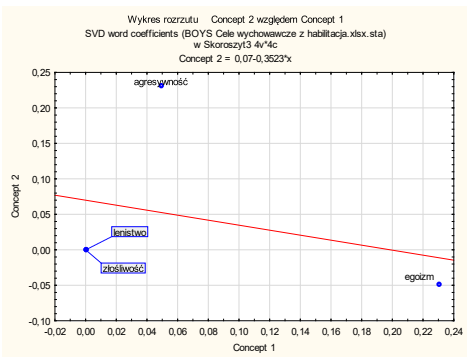
a) Boys: trait 1



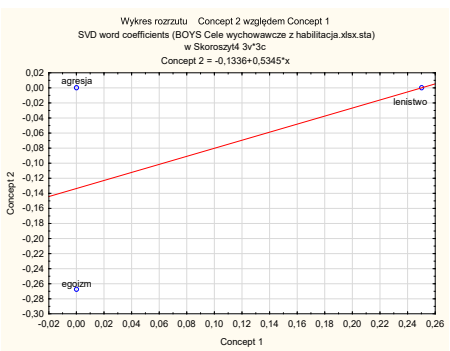
b) Boys: trait 2



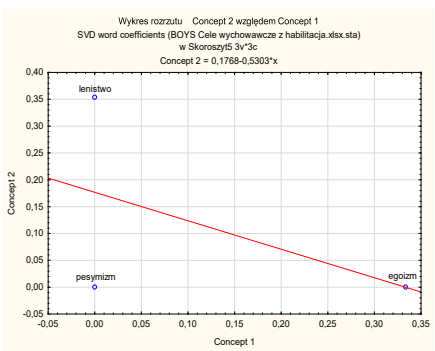
c) Boys: trait 3



d) Boys: trait 4



e) Boys: trait 5



f) Boys: trait 6

Figure 27.2. Results of Principal Component Analysis for parental goals adopted by mothers towards boys

This chapter has demonstrated how text mining algorithms can be effectively used to count words in textual documents, determine their co-occurrence, and analyse the relationships between them. By applying Principal Component Analysis (PCA), it was shown how these techniques can be used to identify hidden patterns and relationships among the traits mentioned by mothers of boys in the context of parental goals.

The word frequency analysis provided a preliminary understanding of the text structure and helped identify the most frequently occurring desirable and undesirable traits. It was shown that traits such as *independence* or *empathy* may form separate clusters, indicating that they are perceived as independent and significant parental goals. Conversely, other traits, such as *assertiveness* and *creativity*, co-occurred, suggesting their complementarity in the eyes of mothers.

Also in the context of undesirable traits such as *aggression*, *egoism*, or *laziness*, PCA allowed for the identification of traits regarded by mothers as particularly harmful and important to avoid. These findings indicate that mothers tend to group certain traits together, considering them particularly detrimental to the development of children's personalities.

The study employed the STATISTICA software, which enabled precise analysis of textual data and the execution of PCA. The application of these analytical tools allows for the identification of both desirable and undesirable traits in the context of child upbringing. These techniques can be widely applied in psychological research, providing valuable insights into parental priorities and values in the upbringing process.

27.2. Transforming Qualitative Data into Quantitative: Methods and Examples

The transformation of unstructured (qualitative) data into structured (quantitative) data is a crucial stage in data analysis using text mining. This chapter presents an example of such a transformation based on the analysis of descriptions of children provided by their mothers.

In the study, mothers were asked to describe their children using various adjectives and phrases that characterised them from different perspectives. Sample descriptions included terms such as *stubborn*, *independent*, *creative*, *cheerful*, *strong*, *brave*, *bright*, and many others. Each mother provided a unique description of her child, which allowed for the collection of a rich and diverse dataset.

Natural language processing algorithms, implemented in the Data Miner module of the STATISTICA software, analysed each description by identifying specific keywords describing the children's traits. These keywords were then entered into a database as new variables, with each variable corresponding to a particular trait, e.g., *ambitious* or *curious*.

To convert qualitative descriptions into quantitative data, numerical values were assigned to each variable based on the presence of a given trait in the description. If a specific word or phrase appeared in a child's description, the algorithm assigned the value 1 to the corresponding variable. The absence of that word resulted in a value of 0. This approach allowed for the creation of a numerical representation of traits, enabling a more objective and quantitative analysis of the data.

Figure 27.3 presents a fragment of the database generated in this process. Each column in the table represents one trait, and each row refers to a child's description provided by the mother. The number "1" indicates the presence of a given trait in the description, while the absence of a value indicates its absence.

The transformation of qualitative into quantitative data, as illustrated in this example, enables the conversion of subjective descriptions into a format suitable for statistical analysis. In psychological, sociological, and market research, such an approach allows for drawing valuable conclusions about the studied phenomena. Thanks to the use of algorithms for the automatic extraction and coding of traits, it becomes possible to obtain precise and objective data that can be subjected to further statistical analysis.

	1 reprezentacja	2 agresywne	3 aktywne	4 ambitne	5 boi	6 bystre	7 chętne	8 ciekawa	9 ciekawe	10 ciekawskie
1	uparte, samodzielne, pomysłowe, wesołe									
2	Silne, odważne, bystre, inteligentne, wrażliwe, impulsywne, chcące rządzić, często się złoszczące i agresywne	1				1				
3	Ciekawe świata, przyrody, radosne, współczujące, dostrzega potrzeby innych, ale też samolubne, nie potrafi się dzielić.								1	
4	inteligentne, ciekawe świata, wrażliwe, gadatliwe, wnikliwe, empatyczne, niezależne, towarzyskie								1	
5	pogodne, szalone, śmiałe, rozgadane, pomysłowe, ruchliwe, silne, skupiające na sobie uwagę									
6	miłe, ciekawe świata, wesołe, inteligentne, z poczuciem humoru								1	
7	wesołe , uparte, ruchliwe ma dużo energii,									
8	wesołe pogodne radosne ciekawe świata zadaje dużo pytań ,ale boi się kontaktów z innymi dziećmi , jest mało pewny siebie , nieśm				1				1	
9	Mądre rozmowne kontaktowe głośno impulsywno energicznie									
10	radosne, pomysłowe, mądre, wrażliwe. Nigdy się nie nudzi, zawsze znajdzie sobie zajęcie. Chętnie zgadza się na różne pomysły, c									
11	urocze, chętne do pomocy, zainteresowane zabawą, samodzielne, ugodowe wobec innych dzieci, nieśmiałe, ale potrafi też być upar						1			
12	żywiołowe, bystre, ciekawe świata, przelotne, rywalizujące, lubi być w centrum zainteresowania, wyraża się artystycznie (spontanic					1			1	
13	Czułe, bardzo inteligentne, odważne, zabawne, piękne, silne									
14	Rozwinięte w pewnych dziedzinach, niezdyscyplinowane, leniwe, uparte, niecierpliwe									
15	Szybkie, nadaktywne, niezależne, nerwowe, inteligentne, pogodne, radosne									
16	Bardzo żywiołowe, ciekawe świata, uparte, chodzące swoimi ścieżkami, zazwyczaj chętny do nauki								1	
17	wesołe, samodzielne, koleżeńskie, otwarte, grzeczne, zabawne									
18	spokojne, współpracujące, wesołe, kreatywne, pewne siebie									
19	Pomocną, pełną życia dziewczynką, wrażliwą.Kocha taniec i muzykę (tak jak mama). Alicja jest ciekawa życia (wszystko od razu							1		
20	Pogodne, uparte, miłe, inteligentne									
21	wybuchowo, labilno, zdolno do kompromisów, skrupulatno, zaciekawiono, uparto, ambitno, płaczliwo, ruchliwo, niepewno siebie,			1						
22	Bardzo zawzięte, głośne, płaczliwe, empatyczne									
23	Oczywiście najcudowniejsze na Świecie ?? jest ciepła, życzliwa, pomocna. Potrzebuje dużo ruchu									
24	temperamentnie, podejmuje szybko decyzje, impulsywne, inteligentne, zdolne									
25	Wesołe, kochane, mądre, dobre, wszędzie jej pełno, potrafi się rozlosic, czasami krzyczy i macha rękami, czasami wymusza									
26	Energiczne, żywiołowe, radosne, kochane									
27	Wesołe,ciekawe świata,pomocne,rozmowne								1	
28	kochane, dobre, miłe, jest moim skarbem									
29	Wesołą, wrażliwą dziewczynką, ciekawą świata istitą , otwartą na nowe doświadczenia. Chętnie się uczy i kojarzy obowiązania mięc									
30	Wyjątkowe dla mnie, czasami uparte, ma swoje zdanie, bardzo ciekawe świata, lubi się wygłupiać i potrafi być empatyczna, ale sta								1	

Figure 27.3. Transformation of child descriptions provided by mothers into numerical form using the Text Mining algorithms of STATISTICA Data Miner

After transforming the child descriptions provided by mothers into numerical data using the Text Mining algorithms of STATISTICA Data Miner, the Generative Pre-trained Transformer 4 (GPT-4) language model, developed by OpenAI, was employed to classify the newly generated variables into two sets: characteristics typical of well-behaved children and characteristics typical of difficult children.

This next step aimed to reduce the number of input variables representing children's traits into a smaller set, thereby facilitating further analysis. Several advanced algorithms were applied in this process. In the first stage, a clustering algorithm was used to group similar traits. Subsequently, feature selection techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were employed to extract the most significant traits.

Using these techniques, the GPT-4 model classified the variables into two sets: traits characteristic of well-behaved children and traits characteristic of difficult children. The solution is presented in Table 27.3.

Table 27.3. Classification of Children's Traits as Well-Behaved or Difficult by the Generative Pre-trained Transformer 4 (GPT-4)

Well-Behaved	Difficult
active – V3, ambitious – V4, bright – V6, willing – V7, curious (f.) – V8, curious (pl.) – V9, inquisitive – V10, wonderful – V11, tender – V12, good – V13, emotional – V14, empathetic (f.) – V15, empathetic (pl.) – V16, energetic – V17, energetic (variant) – V18, well-behaved – V20, intelligent – V24, loves – V25, loving – V26, beloved – V27, sociable – V28, creative – V29, wise (var.) – V32, wise – V33, nice – V34, shy – V39, independent – V40, brave – V41, caring – V42, optimistic – V43, open – V44, beautiful – V45, cheerful – V48, helpful – V49, inventive (var.) – V50, inventive – V51, joyful – V52, lively – V54, self-reliant – V55, strong – V56, sweet – V57, calm – V58, observant – V59, clever – V60, friendly – V61, sincere – V62, happy – V63, outgoing – V64, emotional – V66, conciliatory – V67, charming – V69, smiling (var.) – V70, smiling – V71, talented – V72, attentive – V73, merry (var.) – V74, merry – V75, sensitive (var.) – V76, sensitive – V77, wonderful (var.) – V78, wonderful – V79, funny – V83, resourceful – V84, gifted – V86, lively (var.) – V87, full of energy – V88	aggressive – V2, scared – V5, talkative – V19, impulsive – V22, individualist – V23, lazy – V30, fearful – V31, hyperactive – V35, nervous – V36, impatient – V37, disobedient – V38, tearful – V46, emotional burden – V47, chatty – V53, difficult – V65, stubborn – V68, explosive – V80, outspoken – V81, demanding – V82, jealous – V85

Finally, Generative Pre-trained Transformer 4 (GPT-4) generated code for creating two new variables in STATISTICA. This code could be directly copied into the software, which significantly saved the researcher's time by eliminating the need to manually sum individual trait variables. The code is presented in Table 27.4.

Table 27.4. Code for Creating New Variables in STATISTICA

Well-Behaved	Difficult
well_behaved = V3 + V4 + V6 + V7 + V8 + V9 + V10 + V11 + V12 + V13 + V14 + V15 + V16 + V17 + V18 + V20 + V24 + V25 + V26 + V27 + V28 + V29 + V32 + V33 + V34 + V39 + V40 + V41 + V42 + V43 + V44 + V45 + V48 + V49 + V50 + V51 + V52 + V54 + V55 + V56 + V57 + V58 + V59 + V60 + V61 + V62 + V63 + V64 + V66 + V67 + V69 + V70 + V71 + V72 + V73 + V74 + V75 + V76 + V77 + V78 + V79 + V83 + V84 + V86 + V87 + V88	difficult = V2 + V5 + V19 + V22 + V23 + V30 + V31 + V35 + V36 + V37 + V38 + V46 + V47 + V53 + V65 + V68 + V80 + V81 + V82 + V85

The example presented in this chapter illustrates how Text Mining can be effectively applied to transform subjective descriptions into structures that enable objective quantitative analysis. The transformation of qualitative data into quantitative data through Text Mining is an essential stage in modern data analysis, particularly in the social sciences, market research, and other fields where textual data constitutes a significant source of information.

Text Mining offers numerous benefits in data analysis. First, automating the text processing procedure significantly reduces the time required for analysis. Text Mining algorithms are capable of processing vast amounts of data in a short period, which would be unfeasible with manual analysis. Second, these techniques increase the accuracy of analysis by minimizing the risk of human error, which often

occurs during manual data coding. As a result, more reliable and consistent outcomes can be obtained.

Additionally, Text Mining enables the identification of subtle patterns and relationships that may be difficult for a human to detect. These algorithms can analyse data in a more objective manner, extracting relevant information that might be overlooked using traditional methods.

The use of the Generative Pre-trained Transformer 4 (GPT-4) language model, developed by OpenAI, for data classification brings additional advantages. GPT-4 is capable of conducting large-scale analysis with high precision, reducing subjective bias and ensuring consistency in classification. This model employs advanced machine learning techniques, allowing for more complex and accurate data analysis than traditional manual methods.

Using GPT-4 for trait classification is more efficient than manual analysis, as the model is capable of processing and interpreting data in a more advanced way. This allows for more reliable results to be obtained in a shorter amount of time. Moreover, automating this process eliminates the need for manual data processing, significantly saving the researcher's time and resources. It should also be noted that GPT-4 did not replace the researcher in the entire analysis process, but was deliberately integrated as a tool supporting selected stages, such as code generation. The overall analysis was conducted by the researcher, who guided its direction and interpreted the results.

In summary, the application of Text Mining and advanced algorithms such as GPT-4 in the analysis of qualitative data enables the achievement of more precise, objective, and reliable results. Automating text analysis using these technologies not only increases the accuracy and reliability of the analyses performed but also saves time and resources—an invaluable benefit in the social sciences and other research disciplines.

The data thus prepared may subsequently serve as a foundation for further analyses involving other artificial intelligence algorithms, an example of which will be presented in the next chapter. This makes the analytical process complementary, encompassing the full cycle of data transformation—from unstructured to structured—enabling their more comprehensive use in predictive and exploratory models.

27.3. Integration of Qualitative and Quantitative Data: From Text Mining to Decision Models

This chapter presents an advanced integration of qualitative and quantitative analyses based on unstructured data, which in the previous chapter were transformed into a quantitative form. A previously prepared database, developed using Text Mining techniques, was utilised for further computations through the application of quantitative algorithms. This made it possible to transition to quantitative analysis, demonstrating how variables extracted from qualitative data can be used to explain a quantitative variable—namely, *discrepancy*.

To automate the data analysis process and uncover complex relationships within psychological data, an inductive decision tree algorithm (Classification and Regression Tree, C&RT) was applied. This algorithm is particularly useful, as it enables the identification of significant variables and their influence on selected dependent variables.

Artificial intelligence algorithms allow for the analysis of textual data and the integration of quantitative and qualitative methods on a previously unknown scale (Szymańska, 2017b). They can process very large datasets and do not require specific statistical distributions, making it possible to build diverse models such as decision trees, neural networks, or cluster analyses. The transition from data processed through Text Mining techniques to quantitative analysis using artificial intelligence algorithms is indeed feasible and opens wide-ranging research opportunities.

This chapter will demonstrate how such an approach allows for a more comprehensive understanding of the examined phenomena, enabling better interpretation of complex psychological processes. The integration of these methods not only enhances the analysis but also opens new research perspectives. It allows the inclusion of variables in a single model that would traditionally appear in separate analyses.

The application of Text Mining techniques and advanced algorithms such as decision trees in the analysis of qualitative data allows for more precise, objective, and reliable results. Automating text analysis using these technologies not only increases the accuracy and dependability of conducted analyses but also significantly saves time and resources—an aspect especially valuable in the social sciences and other research domains.

In this chapter, we will focus on explaining the variable *discrepancy* in parental upbringing goals in relation to child traits, using variables derived from the transformation of maternal representations of the child, such as "obedient" and "difficult".

This analysis will provide insight into how mothers' subjective descriptions of their child's traits are linked to their perception of discrepancies in upbringing goals.

The results of the decision tree analysis (C&RT) indicate key variables associated with parental goal discrepancy based on maternal descriptions of the child as "obedient" and "difficult". Below are the most important rules derived from the decision tree analysis, along with node identifiers. The tree structure is illustrated in Figure 27.4, and a detailed representation of the tree is provided in Table 27.5.

- **Rule 1 (Node 2):**

If the level of the "obedient" trait in the child is less than or equal to 23, the parental goal discrepancy is higher, with a mean discrepancy value of 244.47. This suggests that parents of children described by the mother as less "obedient" have more difficulty aligning their upbringing goals.

- **Rule 2 (Node 3):**

If the level of the "obedient" trait in the child is greater than 23, the parental goal discrepancy is lower, with a mean discrepancy value of 131.23. This indicates that parents of children described as more "obedient" by the mother experience fewer difficulties in agreeing on their upbringing goals.

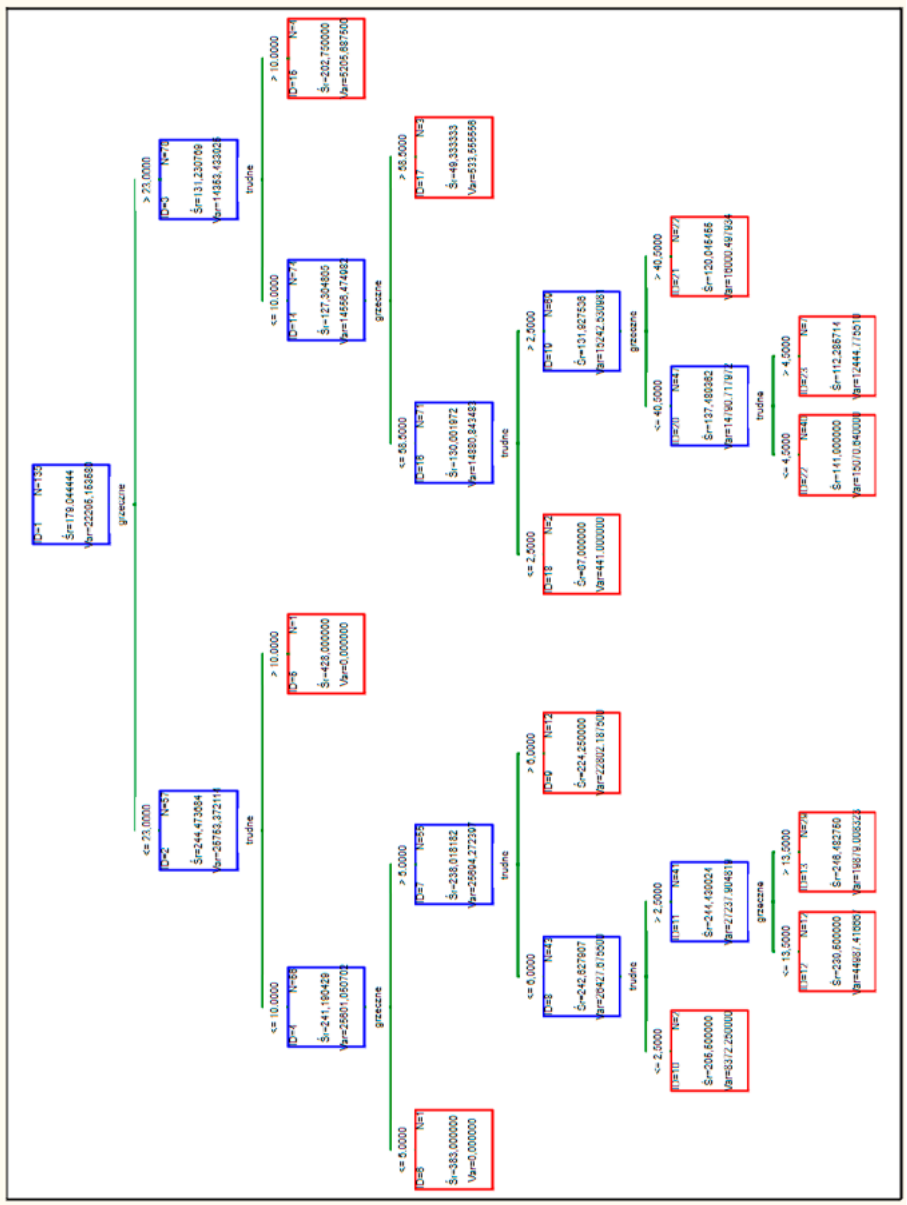


Figure 27.4. Decision Tree Structure for the Variable of Parental Goal Discrepancy
Legend: grzeczne – well-behaved; trudne – difficult; Śr – Mean; Var – Variance; N – Number of cases.

- **Rule 3 (Node 4):**

Among children whose "obedient" level is less than or equal to 23 (Node 2), if the "difficult" trait level is also less than or equal to 10, the parental goal discrepancy is very high, with a mean value of 241.20. This suggests that children who are both less "obedient" and less "difficult" generate the greatest discrepancies in parental upbringing goals.

- **Rule 4 (Node 5):**

Among children whose "obedient" level is less than or equal to 23 (Node 2), if the "difficult" trait level is greater than 10, the parental goal discrepancy is extremely high, with a mean value of 428.00. This means that children who are less "obedient" but more "difficult" generate exceptionally large discrepancies in parental goals.

- **Rule 5 (Node 14):**

In the group of children described as "well-behaved" at a level greater than 23 (node 3), if the level of the "difficult" trait is less than or equal to 10, the discrepancy in parental upbringing goals is even smaller, with a mean discrepancy value of 127.36. This indicates that children who are more "well-behaved" and less "difficult" allow for greater parental agreement in their upbringing goals.

An analysis of the decision tree reveals a general rule that the variables "well-behaved" and "difficult" are significantly related to discrepancies in parental upbringing goals. Higher values of the "well-behaved" trait are associated with lower discrepancies, whereas the level of the "difficult" trait either reduces or increases such discrepancies. Greater discrepancies in parental upbringing goals may co-occur with more challenging child behaviours, evaluated as less "well-behaved".

The rules presented above constitute only a portion of the decision tree analysis but already demonstrate the general pattern revealed by the tree. The use of decision trees enables precise identification of such associations, thereby facilitating a better understanding of the links between subjective descriptions of children and the decisions and attitudes of parents. This, in turn, allows for more objective and detailed findings, which may be utilised in further research and upbringing practices.

Table 27.5. Structure of the Decision Tree for the Variable of Parental Goal Discrepancy

	Left branch	Right branch	Node size	Node mean	Node variance	Split variable	Split value
1	2	3	135	179.0444	22295.15	obedient	23
2	4	5	57	244.4737	25753.37	difficult	10
4	6	7	56	241.1964	25601.05	obedient	5
6	–	–	1	383.0000	0.00	–	–
7	8	9	55	238.6182	25694.27	difficult	6
8	10	11	43	242.6279	26427.68	difficult	2.5
10	–	–	2	205.5000	8372.25	–	–
11	12	13	41	244.4390	27237.90	obedient	13.5
12	–	–	12	239.5000	44987.42	–	–
13	–	–	29	246.4828	19879.01	–	–
9	–	–	12	224.2500	22802.19	–	–
5	–	–	1	428.0000	0.00	–	–
3	14	15	78	131.2308	14353.43	difficult	10
14	16	17	74	127.3649	14556.47	obedient	58.5
16	18	19	71	130.6620	14880.84	difficult	2.5
18	–	–	2	87.0000	441.00	–	–
19	20	21	69	131.9275	15242.53	obedient	40.5
20	22	23	47	137.4894	14790.72	difficult	4.5
22	–	–	40	141.9000	15070.64	–	–
23	–	–	7	112.2857	12444.78	–	–
21	–	–	22	120.0455	16000.50	–	–
17	–	–	3	49.3333	533.56	–	–
15	–	–	4	202.7500	5205.69	–	–

The normality plot of residuals, presented in Figure 27.5, is a key diagnostic tool that allows for the assessment of how closely the residuals of the model—that is, the differences between the predicted and actual values of the dependent variable—follow a normal distribution. This assessment is important because many statistical tests and regression methods assume that residuals are normally distributed.

The points on the plot lie largely close to the red straight line, indicating that the residuals of the model conform to the normal distribution. The closer the points are to this line, the better the model fit. Most of the points in the central part of the plot are located near the red line, suggesting that the distribution of residuals in this section is normal. At the extreme ends of the plot (left and right), certain deviations from the line can be observed, particularly on the right-hand side. These deviations may indicate the presence of outliers or some degree of asymmetry in the data.

The points are relatively symmetrically distributed around the line, which is a positive sign. The absence of noticeable clustering on one side of the line suggests that the distribution of residuals is symmetric. Observations at the extreme ends of the plot that deviate significantly from the line may indicate the presence of

outliers. In our case, several points on the right side show greater deviations, suggesting that certain extreme values may be present in the data.

Most points lie close to the straight line, indicating that the model's residuals are largely consistent with a normal distribution. This is a positive sign and indicates a good model fit to the data. A few points at the far ends of the plot show larger deviations from the line, particularly on the right-hand side. This points to the possibility of outliers that may influence the results of the analysis.

In summary, the normality plot of residuals for our model shows that the residuals are largely consistent with a normal distribution, which is favourable. Nevertheless, certain deviations at the extreme ends suggest that attention should be paid to potential outliers and asymmetry in the data.

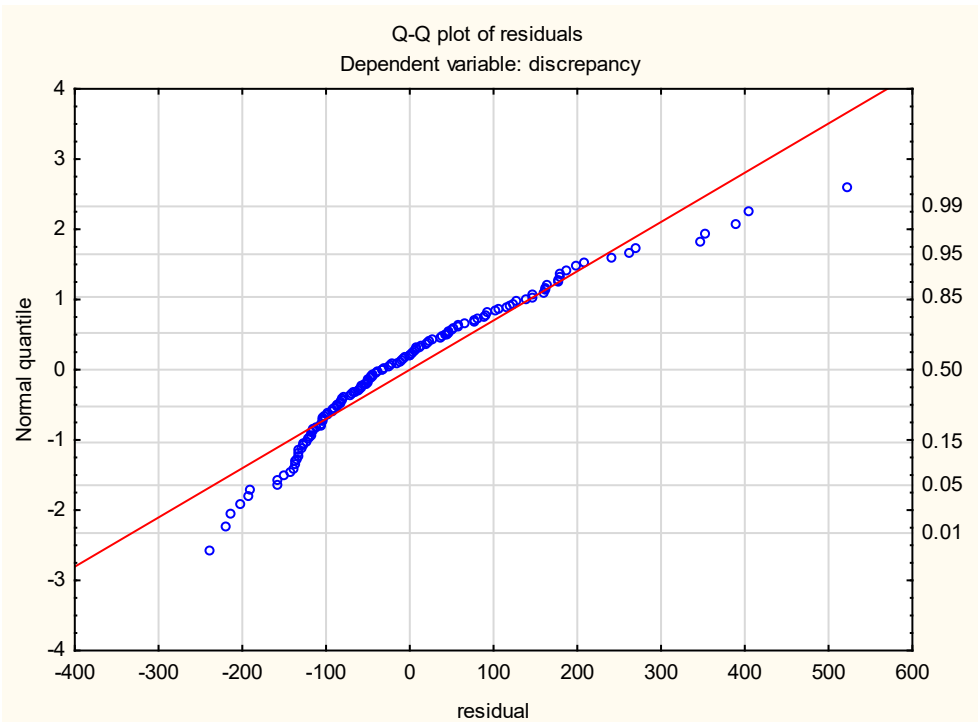


Figure 27.5. Normality Plot of Residuals

The parallel coordinates plot presented in Figure 27.6 illustrates the relationships between the variables *well-behaved*, *difficult*, and the dependent variable *discrepancy* (i.e., parental goal discrepancy). This is a visualisation tool that allows one to observe how the values of these variables influence the discrepancy in parental upbringing goals.

Each vertical axis on the plot represents one of the variables: *well-behaved*, *difficult*, and *discrepancy*. The variable values are plotted along these axes, and the lines connecting points across the axes show how specific values of one variable are associated with values of the other variables in the dataset. The pattern of these lines

makes it possible to identify trends and relationships among the variables.

On the left edge of the plot, we observe the values of the *well-behaved* variable, which range from 1 to 90. Higher values of this variable are associated with lower values of *discrepancy*, suggesting a smaller discrepancy in parental upbringing goals. In turn, the values of the *difficult* variable, shown on the central axis, range from 1 to 13. Higher values of this variable are often linked to higher *discrepancy* values, indicating a greater divergence in parental goals.

On the right edge of the plot are the values of the *discrepancy* variable, ranging from 1 to 762. Lines connecting lower *well-behaved* values and higher *difficult* values to higher *discrepancy* values indicate significant associations between these variables and the parental goal discrepancy.

The parallel coordinates plot demonstrates a clear relationship between the *well--behaved* and *difficult* variables and the *discrepancy* variable. Higher values of the *well-behaved* trait are associated with smaller discrepancies in parental goals, while higher values of the *difficult* trait are linked to greater discrepancies.

The parallel coordinates plot provides valuable insights into the relationships among variables. It is a useful tool for visualising associations that may be difficult to detect in traditional data tables. Thanks to this plot, we can gain a better understanding of how various child traits, as described by the mother, influence the discrepancy in parental upbringing goals.

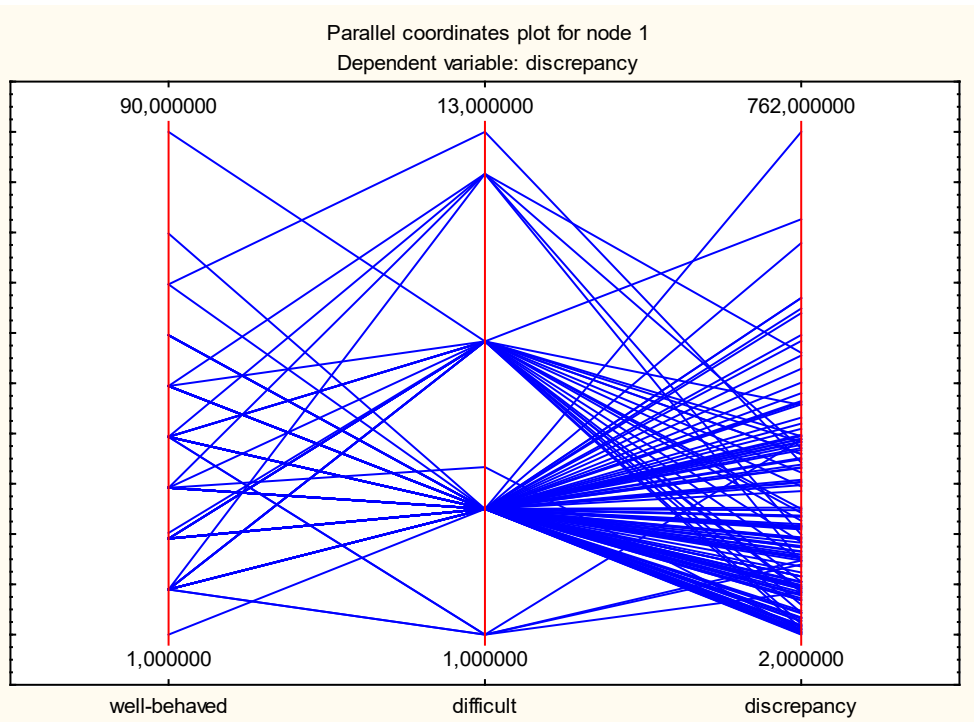


Figure 27.6. Parallel Coordinates Plot for Node 1

In the presented analysis, we followed the entire process of data transformation—from unstructured parental interview descriptions, through their organisation using Text Mining algorithms, to classification and further quantitative analysis employing artificial intelligence.

First, the Text Mining algorithm assigned traits to each participant based on the descriptors found in their narrative, thereby generating a structured database. Next, the GPT-4 model classified these traits, assigning them to two categories—those characteristic of *well-behaved* and *difficult* children. The subsequent step involved GPT-4 generating code that enabled the automatic creation of new variables in STATISTICA, summing the traits belonging to each of the defined categories.

After transitioning from raw descriptions to measurable variables, the next stage of analysis was conducted—application of a decision tree, another artificial intelligence algorithm, which allowed for the identification of key relationships between the newly derived variables and the dependent variable. This enabled a complete transformation from unstructured to structured data, facilitating comprehensive quantitative analysis.

The final results, presented in the plots, demonstrate how effective the integration of qualitative and quantitative methods can be in psychological research through the use of artificial intelligence algorithms. This process opens new possibilities in data analysis, allowing for more precise identification of patterns and relationships that were previously difficult to detect.

CHAPTER 28

Web Crawling Techniques and Online Data Analysis

Following a comprehensive analysis in which qualitative data were transformed into quantitative form and artificial intelligence algorithms were employed to uncover patterns, it is time for the next step. Until now, the research was based on ready-made data derived from interviews, but what if the researcher wishes to acquire data directly from the Internet? In today's world, where a vast amount of information is available online, the ability to efficiently search, retrieve, and process web-based data becomes crucial.

This chapter presents methods for automatically collecting data from the Internet using web crawling and web scraping techniques. Tools enabling the systematic exploration of websites will be introduced, along with methods for extracting relevant information and analysing the collected data. As a result, the reader will acquire the skills necessary to obtain large datasets in an automated manner, thereby opening new opportunities for scientific research and data analytics. This means that the reader will not only learn how to analyse data but also how to acquire it—since the true power of analysis lies not only in the methods but also in the ability to access valuable information.

Web crawling and online data analysis are key techniques in today's data-driven world. Web crawling—the indexing of websites—involves the automated scanning of web pages to gather and organise information (Elder et al., 2012). Web scraping, by contrast, focuses on extracting specific data from already visited websites. These two methods are often used together: first, to identify relevant resources (crawling), and then to retrieve the required information (scraping) (Glez-Peña et al., 2013). Online data analysis, in turn, is concerned with processing and interpreting the

collected data to derive meaningful insights (Zhang & Segall, 2008). This chapter will discuss fundamental web crawling and web scraping techniques, the tools available for these tasks, as well as approaches to analysing the collected data.

28.1. Methods and Tools for Web Searching

Web crawling is widely used by search engines such as Google, Bing, and Yahoo to build their indexes. This process involves "bots" or "spiders" that crawl web pages, collecting data and links that lead to other pages. As a result, search engines can respond quickly and efficiently to user queries, providing the most relevant results (Nisbet et al., 2009). Examples of tools used for web crawling include Scrapy, Heritrix, and Apache Nutch—a powerful open-source framework for building search engines.

Web scraping, by contrast, is a more targeted technique aimed at extracting specific data from selected websites. Various tools and programming libraries are used for this purpose, such as BeautifulSoup, Selenium, and Puppeteer. BeautifulSoup, for instance, is a Python library that allows for efficient parsing of HTML and XML documents, enabling the extraction of relevant information. Selenium is a browser automation tool used not only for testing web applications but also for complex scraping operations that require interaction with dynamically generated content (Yuan, 2023). In practice, web scraping can be applied to a range of purposes—from monitoring product prices in online stores and analysing sentiment on social media to collecting data for scientific research.

Online data analysis, the next step after data collection, involves processing and interpreting the acquired data. It includes a variety of analytical techniques, such as statistical analysis, data mining, and machine learning methods. One example is the use of tools like Google Analytics to monitor website traffic, which allows for optimisation of content and marketing strategies.

In summary, web crawling and web scraping are tools that enable efficient searching and analysis of web resources. Their applications are broad, encompassing both commercial and scientific projects. The practical use of these techniques, however, requires familiarity with appropriate tools and technologies, as well as an awareness of the ethical and legal considerations related to web data processing.

28.2. Practical Workshop and Tutorials

This chapter is dedicated to two key techniques for acquiring data from the Internet using the STATISTICA software: Web Crawling and Web Scraping.

The aim of this chapter is to demonstrate how to perform Web Crawling and Web Scraping in STATISTICA in order to collect and analyse web-based data. The examples provided here teach how to configure these processes, as well as

how to transform and interpret the collected data—opening up broad research and analytical opportunities.

28.2.1. Web Crawling in STATISTICA

To conduct Web Crawling, one can use STATISTICA, which enables not only the scanning of websites but also the immediate analysis of the collected data using Text Mining algorithms. To begin, select:

- **Data Mining > Web Crawling, Document Retrieval**

A window will appear containing various options and functions necessary for performing web crawling and analysing the retrieved data (Figure 28.1).

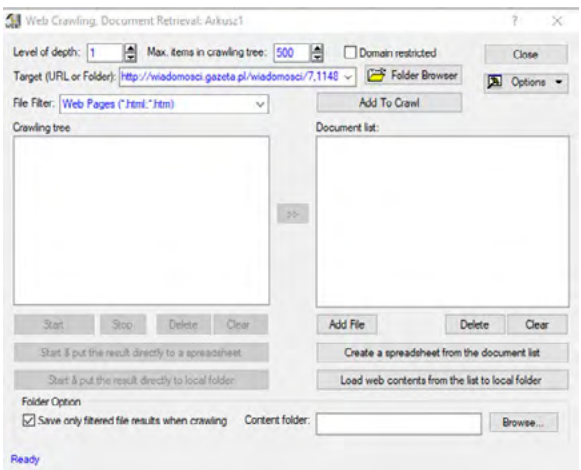


Figure 28.1. Configuration window for web crawling and document retrieval in STATISTICA

Level of depth allows the user to define the depth of website traversal. Setting this value to 1 means that only the main pages will be crawled, without navigating into subpages. This is useful when it is necessary to quickly collect basic information without delving into the site's full structure.

Max. items in crawling tree defines the maximum number of elements that may be crawled. The default value is 500, which should suffice for most applications. This option allows the user to control the volume of data and the resources consumed during the crawling process.

Domain restricted ogranicza przeszukiwanie do jednej domeny, zapewniając, że crawler nie przejdzie na inne strony. Jest to szczególnie ważne, gdy konieczne jest skupienie się na danych z konkretnej witryny.

Domain restricted limits the crawling process to a single domain, ensuring that the crawler does not navigate to external websites. This is especially important when the focus is on data from a specific site.

The field *Target* (URL or Folder) is used to enter the URL of the website to be crawled. This defines the source from which data will be collected. Providing the correct URL is essential for the crawler to start in the right place.

File Filter enables the selection of file types to be crawled. For example, selecting "Web Pages (*.html;*.htm)" will limit crawling to HTML-format web pages, which is the most typical setting for web crawling.

The *Add To Crawl* button adds the entered URL to the crawling list. Once added, the address will appear in the crawling tree, which represents the structure and scope of the crawl.

The *crawling tree* presents a visual representation of the crawl structure, displaying the added URLs. This helps to better understand and manage the crawling process.

The *document list* displays the documents collected during the crawling session. Once crawling is complete, all retrieved documents will appear here, ready for further analysis or saving.

The *Start, Stop, Delete, and Clear* buttons are used to manage the crawling process. Start initiates crawling, Stop interrupts it, Delete removes selected items from the list, and Clear clears the entire list.

The *Folder Option* contains settings related to saving results. Selecting the option save only filtered file results when crawling ensures that only filtered results will be saved, which is useful when one is interested in specific types of data. The Content folder field allows the user to select the location where the collected data will be saved. Clicking the Browse button makes it possible to specify the destination folder on the local drive.

The button *create* a spreadsheet from the document list enables the creation of a spreadsheet from the list of collected documents, allowing for the immediate transformation of the collected data into a more analysis-friendly format.

The final function, *load web contents* from the list to local folders, makes it possible to load the contents of the collected web pages into the selected local folder. This ensures easy access to the collected data, which can then be reviewed directly from the local drive.

These functions enable full management of the crawling and data saving process, allowing for further analysis within the STATISTICA software.

To begin web crawling using STATISTICA, the first step is to select a topic of interest. In this case, the topic is "Chat GPT". In the program window, the appropriate URL of the Wikipedia page containing information on this topic is entered. In the Target (*URL or Folder*) field, one enters https://en.wikipedia.org/wiki/ChatGPT, as shown in the attached figure (Figure 28.2, point 1).

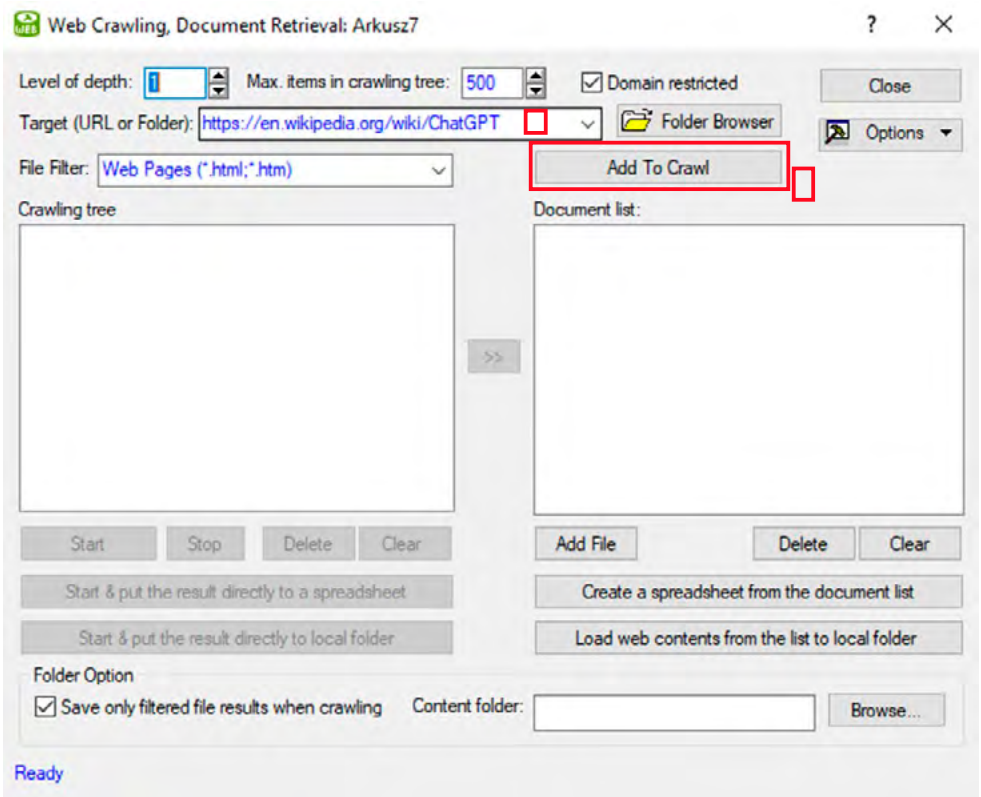


Figure 28.2. Configuration window for web crawling and document retrieval in STATISTICA, steps 1, 2, and 3

We want the crawler to search for information about ChatGPT exclusively within Wikipedia. Therefore, the *Domain restricted* option is selected to limit the crawling process to a single domain (Figure 28.2, point 2). Since we do not want to crawl subpages, the *Level of depth* remains set to 1. We click *Add To Crawl*, which transfers the selected page to the *Crawling Tree* window (Figure 28.2, point 3). The maximum number of items to crawl is set to 500, meaning that the crawler will retrieve up to 500 pages.

After adding the URL to the crawling list and ensuring that all relevant options are selected, we can begin the web crawling process by clicking the Start button (Figure 28.3, point 4).

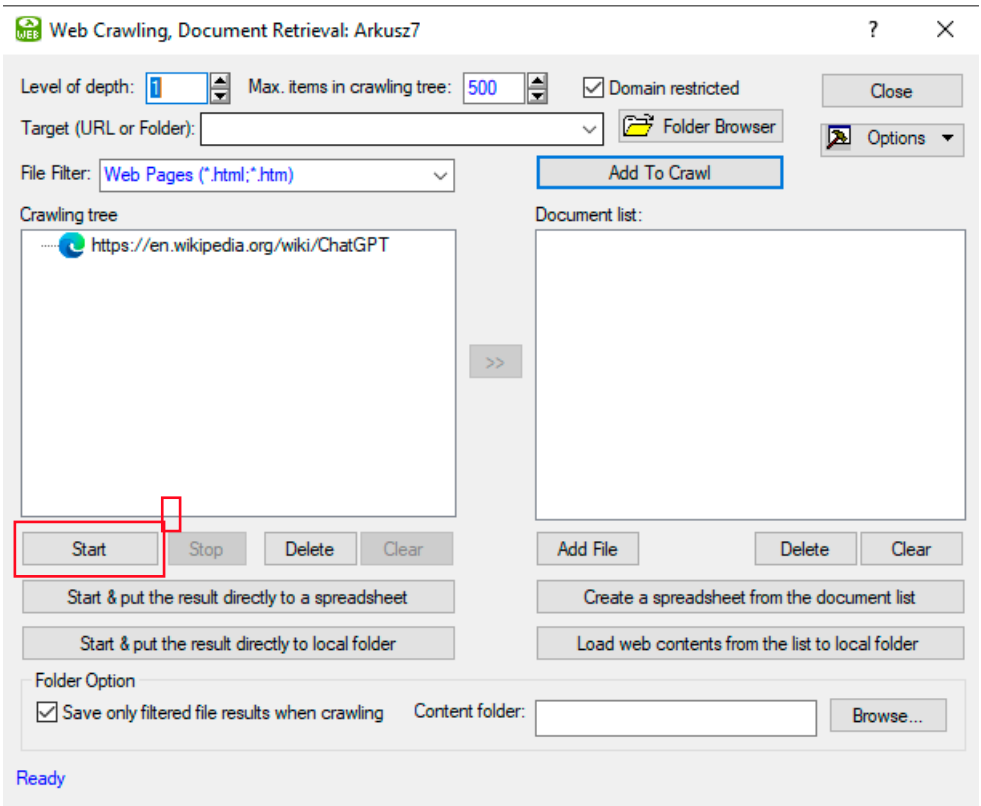


Figure 28.3. Configuration window for web crawling and document retrieval in STATISTICA, step 4

During the crawling process, additional links will appear in the Crawling Tree window, as shown in Figure 28.4, point 5. To stop the crawling process, one can press the STOP button (Figure 28.4, point 6). If the process is not interrupted manually, web crawling will end automatically once it reaches 500 items or when the topic has been exhausted online.

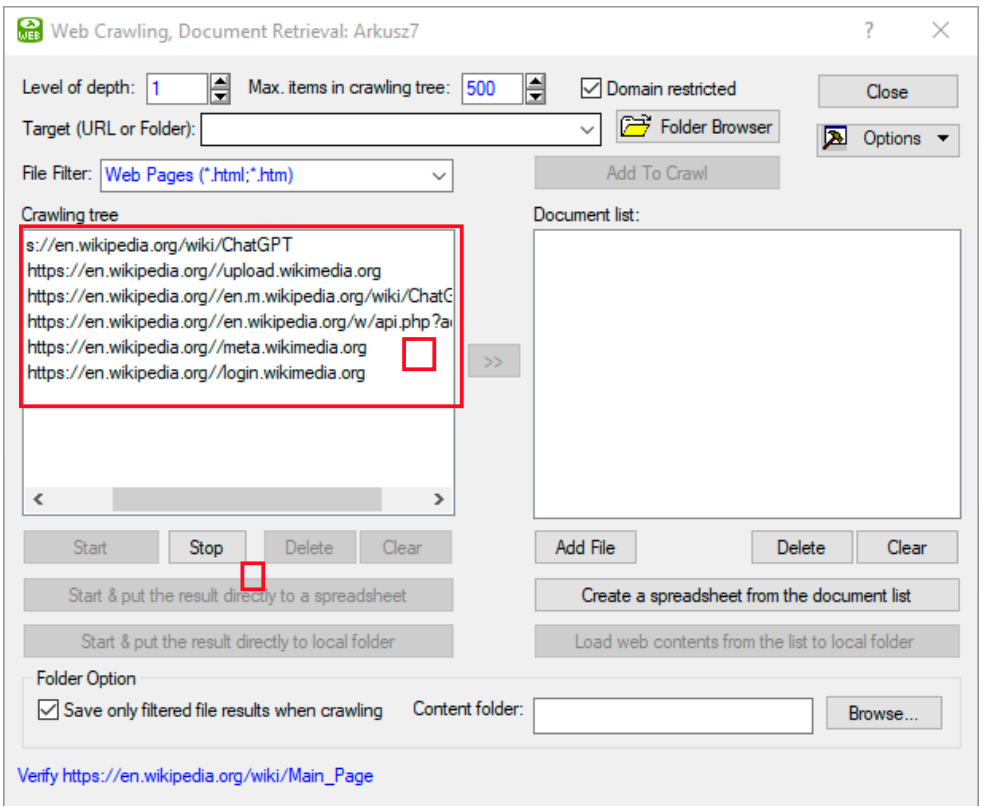


Figure 28.4. Configuration window for web crawling and document retrieval in STATISTICA, steps 5 and 6

After the crawling process is complete, the web crawling session ends, and the next step is to save the results. To do this, click the button *Start & put the results directly to a spreadsheet*. This will save the crawled pages in a new sheet tab of the spreadsheet (Figure 28.5, point 7).

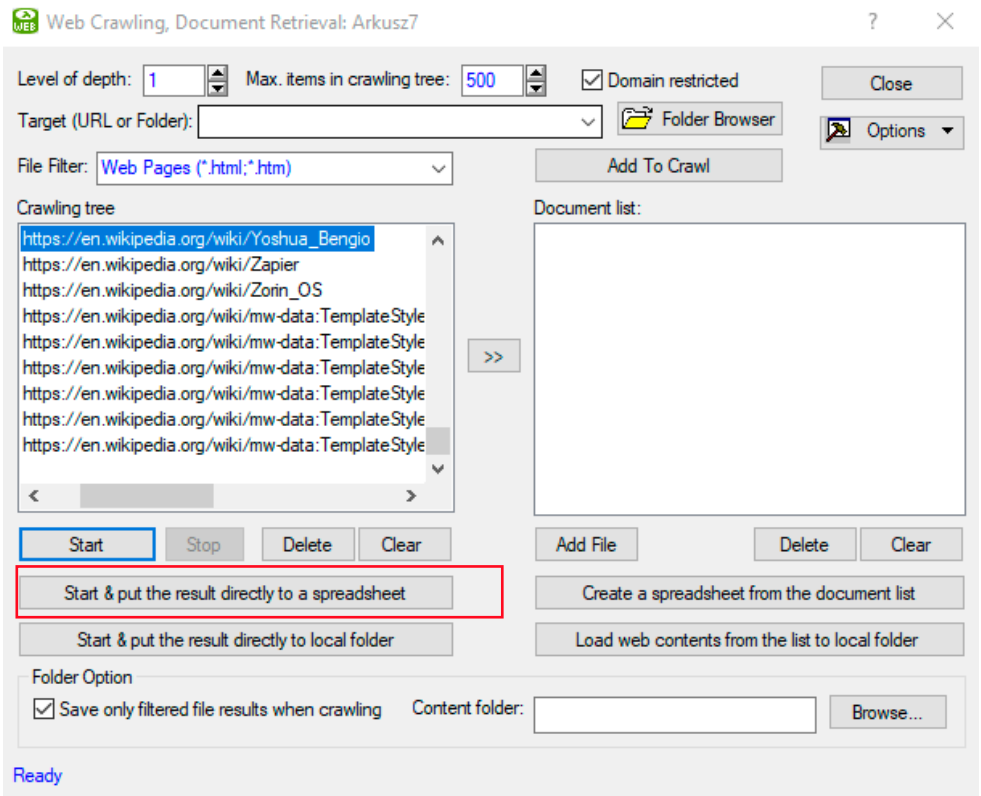
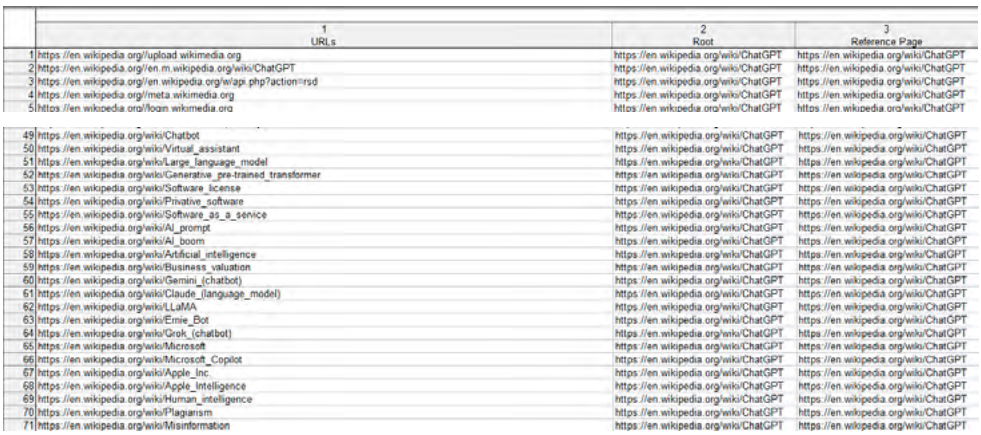


Figure 28.5. Configuration window for web crawling and document retrieval in STATISTICA, step 7

In the new sheet tab, all links to pages related to ChatGPT will be saved. The reviewed links reveal many related topics, such as chatbots, virtual assistants, Generative Pre-trained Transformer, human intelligence, plagiarism, and others. These include not only those directly concerning ChatGPT, but also those thematically associated with it (see Figure 28.6).



	1 URLs	2 Root	3 Reference Page
1	https://en.wikipedia.org//upload.wikimedia.org	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
2	https://en.wikipedia.org//en.m.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
3	https://en.wikipedia.org//en.wikipedia.org/w/api.php?action=rsd	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
4	https://en.wikipedia.org//meta.wikimedia.org	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
5	https://en.wikipedia.org//login.wikimedia.org	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
49	https://en.wikipedia.org/wiki/Chatbot	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
50	https://en.wikipedia.org/wiki/Virtual_assistant	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
51	https://en.wikipedia.org/wiki/Large_language_model	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
52	https://en.wikipedia.org/wiki/Generative_pre-trained_transformer	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
53	https://en.wikipedia.org/wiki/Software_license	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
54	https://en.wikipedia.org/wiki/Privative_software	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
55	https://en.wikipedia.org/wiki/Software_as_a_service	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
56	https://en.wikipedia.org/wiki/AI_prompt	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
57	https://en.wikipedia.org/wiki/AI_boom	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
58	https://en.wikipedia.org/wiki/Artificial_intelligence	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
59	https://en.wikipedia.org/wiki/Business_valuation	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
60	https://en.wikipedia.org/wiki/Gemini_(chatbot)	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
61	https://en.wikipedia.org/wiki/Claude_(language_model)	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
62	https://en.wikipedia.org/wiki/LLaMA	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
63	https://en.wikipedia.org/wiki/Ernie_Bot	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
64	https://en.wikipedia.org/wiki/Grok_(chatbot)	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
65	https://en.wikipedia.org/wiki/Microsoft	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
66	https://en.wikipedia.org/wiki/Microsoft_Copilot	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
67	https://en.wikipedia.org/wiki/Apple_Inc.	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
68	https://en.wikipedia.org/wiki/Apple_Intelligence	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
69	https://en.wikipedia.org/wiki/Human_intelligence	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
70	https://en.wikipedia.org/wiki/Plagiarism	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT
71	https://en.wikipedia.org/wiki/Misinformation	https://en.wikipedia.org/wiki/ChatGPT	https://en.wikipedia.org/wiki/ChatGPT

Figure 28.6. List of URLs for analysis in STATISTICA

The interconnection of these topics is particularly interesting, as even at this stage the data can be used to perform a Text Mining analysis to determine which subjects are most frequently associated with ChatGPT. Text Mining makes it possible to identify patterns and relationships among various concepts, which may lead to new insights and a deeper understanding of how ChatGPT is perceived and applied in different contexts. For instance, the analysis may reveal that ChatGPT is often mentioned in the context of education, medicine, or marketing—suggesting its broad applicability in these domains.

To save the results, one may also use the option *Start & put the results directly to local folder*, which allows the data to be saved directly to a selected local folder. This ensures easy access to the crawled data at any time and enables further analysis or processing using other analytical tools.

These functions make it possible not only to collect data, but also to analyse it immediately, facilitating the rapid extraction of useful insights and understanding the context in which ChatGPT operates, as well as identifying related topics. Such analysis may be highly valuable for researchers, data analysts, and companies seeking to better understand how their technologies are perceived and what potential applications they may have.

TEXT MINING

To better understand the topics associated with ChatGPT, an analysis is conducted on data obtained through web crawling. Such an analysis provides a general overview

of the subject, which is particularly useful when detailed knowledge about the topic is lacking.

After completing the web crawling procedure and saving the results in a spreadsheet, the first stage of analysis can begin. To initiate the analysis, go to the *Data* menu and select *Input Spreadsheet*. Then, choose the *Text & Document Mining* option to begin analysing the texts contained in the collected links.

The STATISTICA window shown in Figure 28.7 is associated with the text exploration module (Text Mining). In the upper part of the window are tabs for configuring various text analysis options, such as Filters, Characters, Delimiters, Defaults, Project, Quick, Advanced, and Words. These tabs allow the user to fine-tune the parameters of text mining, from basic settings to more advanced configurations.

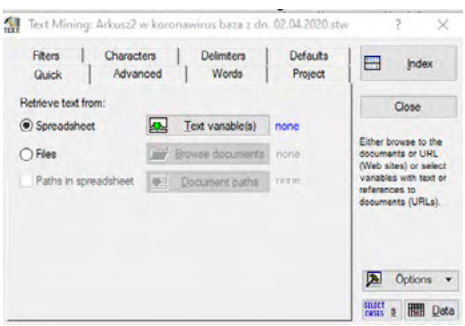


Figure 28.7. Source selection window for text analysis in the Text Mining module in STATISTICA

The central part of the window is dedicated to selecting the source of text for analysis. The user may choose from three options: spreadsheet, files, or paths in spreadsheet. In this case, the option "Spreadsheet" is selected, meaning that the text will be retrieved from a spreadsheet. Below, there is a clickable field labelled "Text variable(s)", currently marked as "none", indicating that no text variable has yet been selected. The "Files" and "Paths in spreadsheet" options are currently inactive, meaning that the corresponding buttons for browsing files and file paths are greyed out and cannot be used.

On the right-hand side, there are action buttons, including "Index", which indexes the selected text, "Close", which closes the window, and "Options" for additional text mining settings. Additionally, there are options for "Select Cases", allowing the selection of specific cases for analysis, and "Data", which most likely enables the viewing and management of data.

At the bottom of the window, there is a message instructing users to browse documents or URLs (web pages), or to select variables containing text or document references (URLs).

In summary, this window in STATISTICA is essential for configuring the initial parameters of text mining, enabling the selection of appropriate data sources and variables to be analysed. This allows the user to effectively manage the text mining

process by tailoring the settings to the specific needs of the analysis.

After clicking "Text variables", a window appears in which the first variable is URL. We select it and then click "OK" (Figure 28.8).

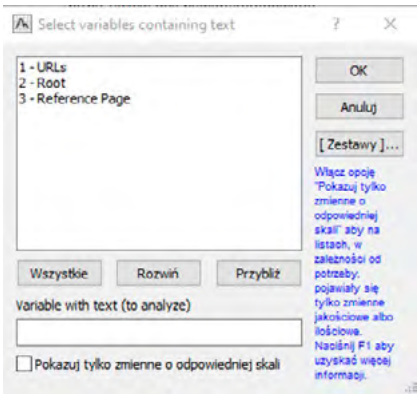


Figure 28.8. Window for selecting text variables for analysis in STATISTICA

In the settings tab labelled "Advanced", the option "Keep unselected words in database for browsing" is selected. This function ensures that unselected words are retained in the database, allowing them to be reviewed and analysed at a later stage. This enables more comprehensive and flexible management of all words in the text, even those not initially chosen for direct analysis (Figure 28.9).

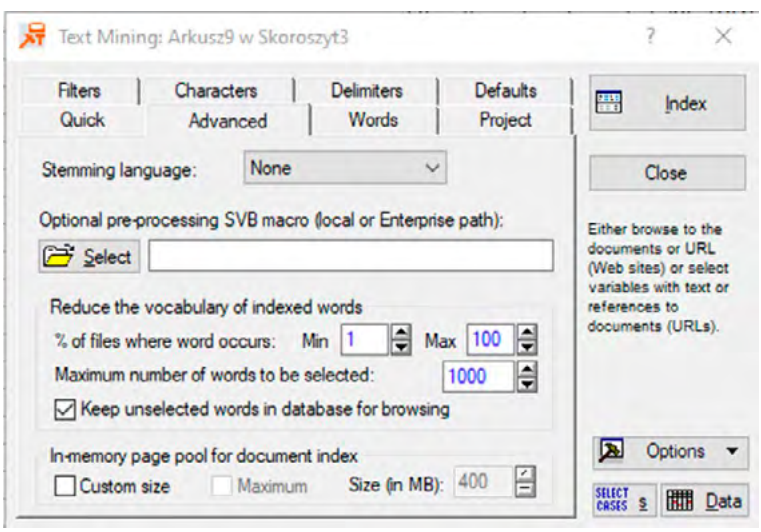


Figure 28.9. Advanced text indexing settings in the Text Mining module in STATISTICA

After completing the previous steps, we return to the main text mining window in STATISTICA, which was shown earlier. In this window, after confirming that all settings are correct, we click the "Index" button. This prompts the program to begin indexing the selected text, which is a necessary step for further textual data analysis. Indexing enables faster and more efficient processing and searching of the text, facilitating the subsequent stages of analysis (Figure 28.7).

A list of all words is generated, as shown in Figure 28.10.

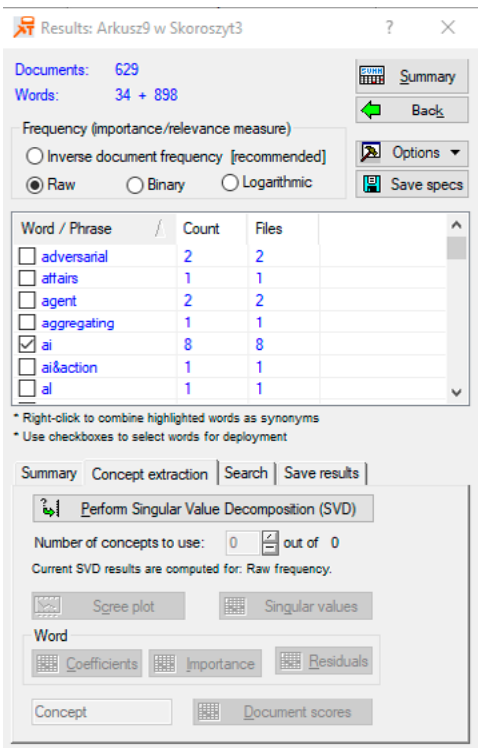


Figure 28.10. Word frequency for the ChatGPT analysis

Next, a Principal Component Analysis was conducted in order to identify word clusters. This analysis also helps determine which words are associated with ChatGPT (Figure 28.11).

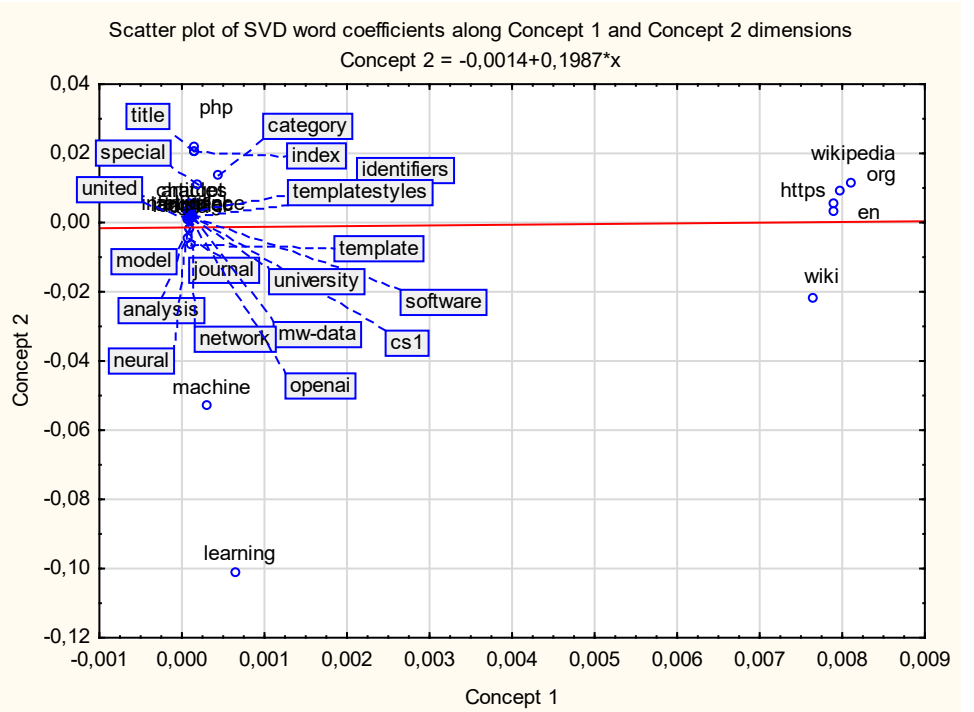


Figure 28.11. Principal Component Analysis for the ChatGPT analysis

This chart shows that even without any prior knowledge of what ChatGPT is, one can quickly identify numerous associations with machine learning, artificial intelligence, neural networks, models, and analysis. The chart displays words such as "neural", "machine", "network", "learning", "model", and "analysis", which clearly indicate connections with advanced technologies. Thus, we are dealing with a system that is a model based on artificial intelligence and machine learning. Such a chart makes it possible to understand that ChatGPT is a tool that utilises the latest advancements in AI, making it a highly sophisticated natural language processing system.

Of course, the researcher may also manually highlight key words—not for general orientation, but to present a broader picture—especially when possessing in-depth knowledge of the topic under investigation. The author, for instance, highlighted two words: "transformer" and "transformers", as it is difficult to discuss ChatGPT without referencing transformers (Figure 28.12).

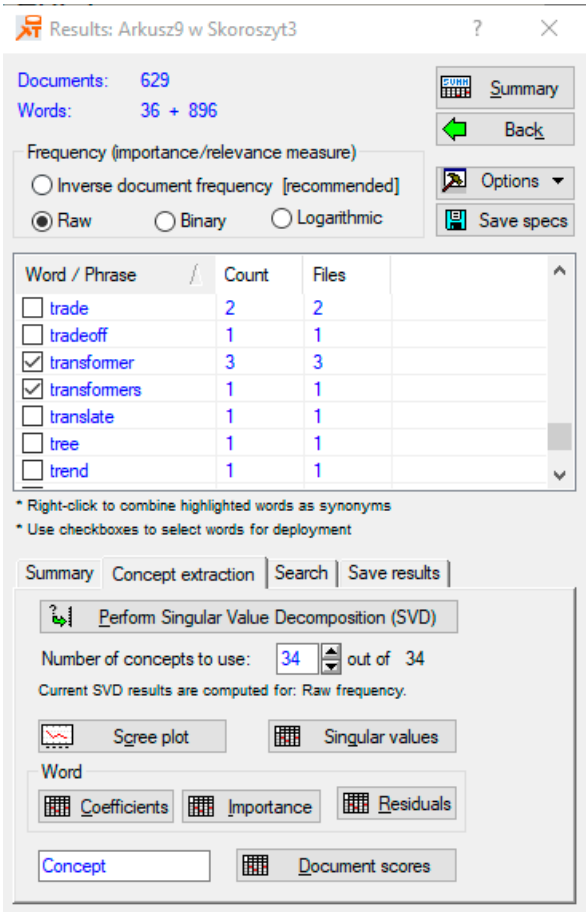


Figure 28.12. Word frequency for the ChatGPT analysis including the words "transformer" and "transformers"

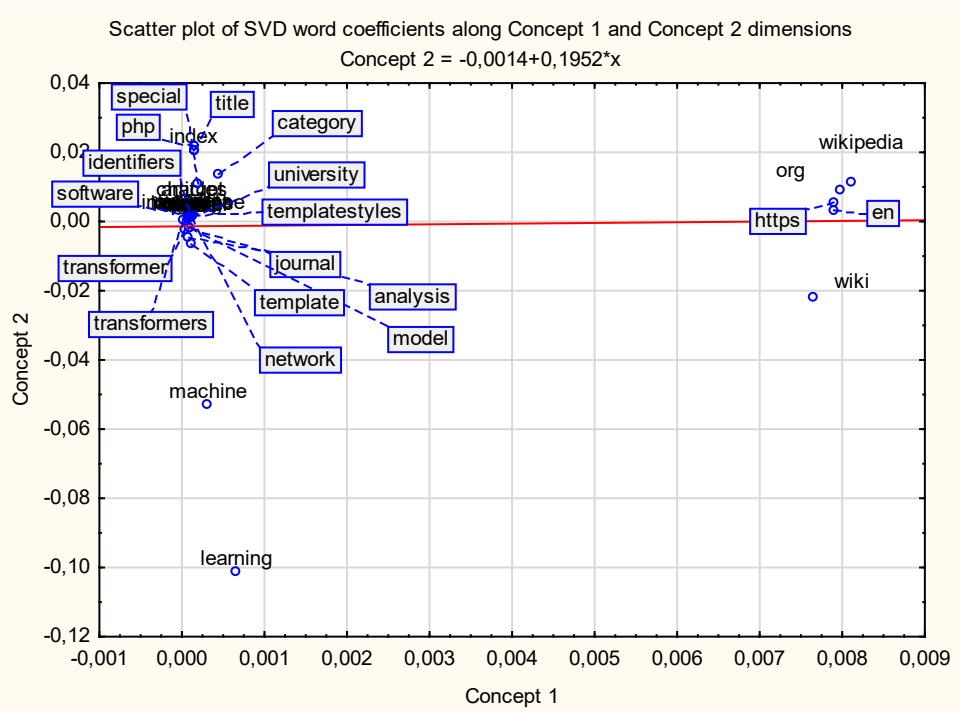


Figure 28.13. Principal Component Analysis including the words "transformer" and "transformers"

These two words therefore appear in the subsequent Principal Component Analysis, highlighting their relevance to the subject under investigation (Figure 28.13). Web crawling thus plays a key role in preparing material for more detailed analyses enabled by web scraping. Web crawling provides a broad overview of a given issue by collecting data from multiple sources, allowing for a general understanding of the subject. Then, through more targeted extraction and analysis, web scraping facilitates deeper insight and more precise processing of the collected data. This combined approach enables researchers to examine the topic from a comprehensive and multidimensional perspective.

28.2.2. Web Scraping in STATISTICA

Another highly interesting application of the STATISTICA software is its capability to perform web scraping. This process follows the preceding stage of web crawling. After scanning the web and collecting data, web scraping allows for the extraction of information from websites.

Web scraping involves the automated retrieval of data from websites. With this functionality, STATISTICA enables the collection of information from various web

pages and the saving of this information in an organised manner—an invaluable feature for analysing large volumes of data.

The web scraping process in STATISTICA begins with selecting all the links to the web pages from which data is to be retrieved. These links are transferred to the "Document List" window using an arrow within the program. The software then scans the selected websites, extracting various types of content such as text, images, tables, headers, and other elements embedded in the pages.

The extracted data are subsequently saved in a defined format, such as CSV files, Excel spreadsheets, databases, or other structured formats that facilitate further analysis and processing. As a result, STATISTICA enables users to efficiently collect and organise large amounts of web data.

Web scraping allows for the retrieval of various types of content. Textual elements such as articles, descriptions, and comments can be extracted, as well as images including photographs, graphics, and charts. Additionally, tabular data such as results, statistics, and product prices, along with headers and metadata such as page titles, meta descriptions, and tags, can also be collected and analysed.

However, certain limitations must be taken into account. Not all websites permit web scraping; some may actively block such activities. It is also essential to ensure that the retrieval of data complies with applicable laws and the terms of service of the websites in question.

Web scraping in STATISTICA enables comprehensive data extraction from the Internet, which can then be analysed and processed within the program—providing vast research and analytical possibilities. This tool allows users to effectively manage and analyse large datasets, which is extremely beneficial across numerous fields of research.

To carry out this process, we first select all the links to the web pages we wish to save, and then transfer them using the arrow to the Document List window (Figure 28.14, steps 1 and 2).

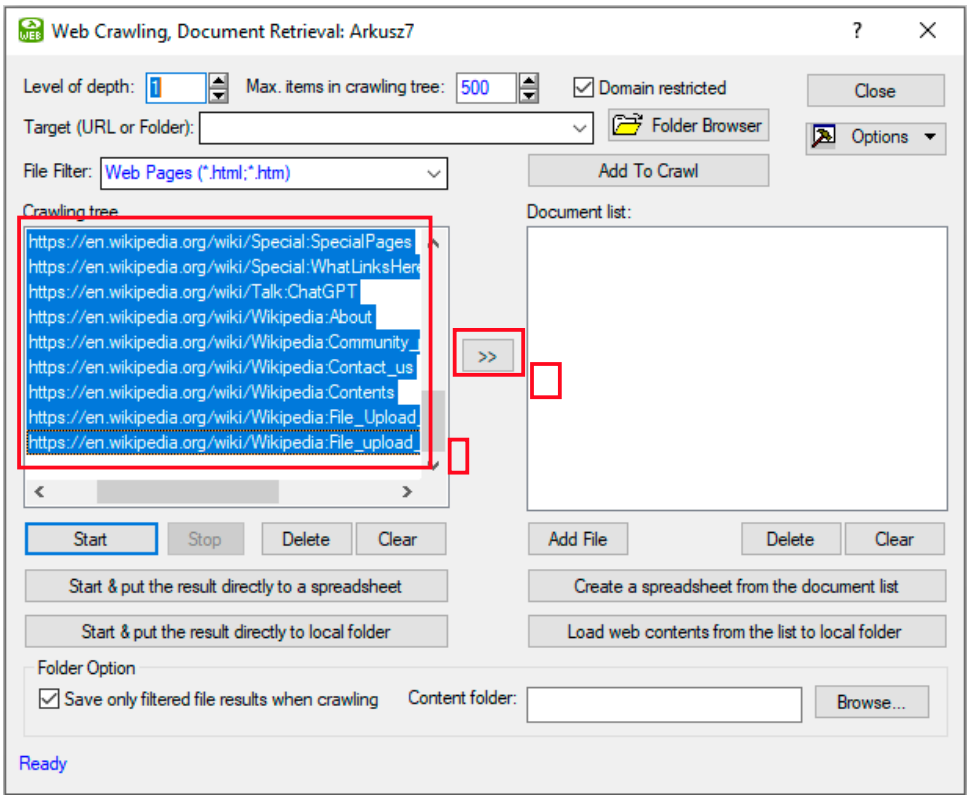


Figure 28.14. Web scraping procedure, steps 1 and 2

Next, we select the links in the "Document List" window that we wish to process and click "Load web contents from the list to local folder". At this point, the process of downloading the content of the selected web pages and saving it to the specified local folder begins. This process enables the automatic transformation of online data into structured files on the local machine, allowing for their subsequent analysis and processing without the need for continuous Internet access. With this functionality, STATISTICA allows users to efficiently and conveniently manage large collections of web data, ensuring access to the information at any time (Figure 28.15, steps 3 and 4).

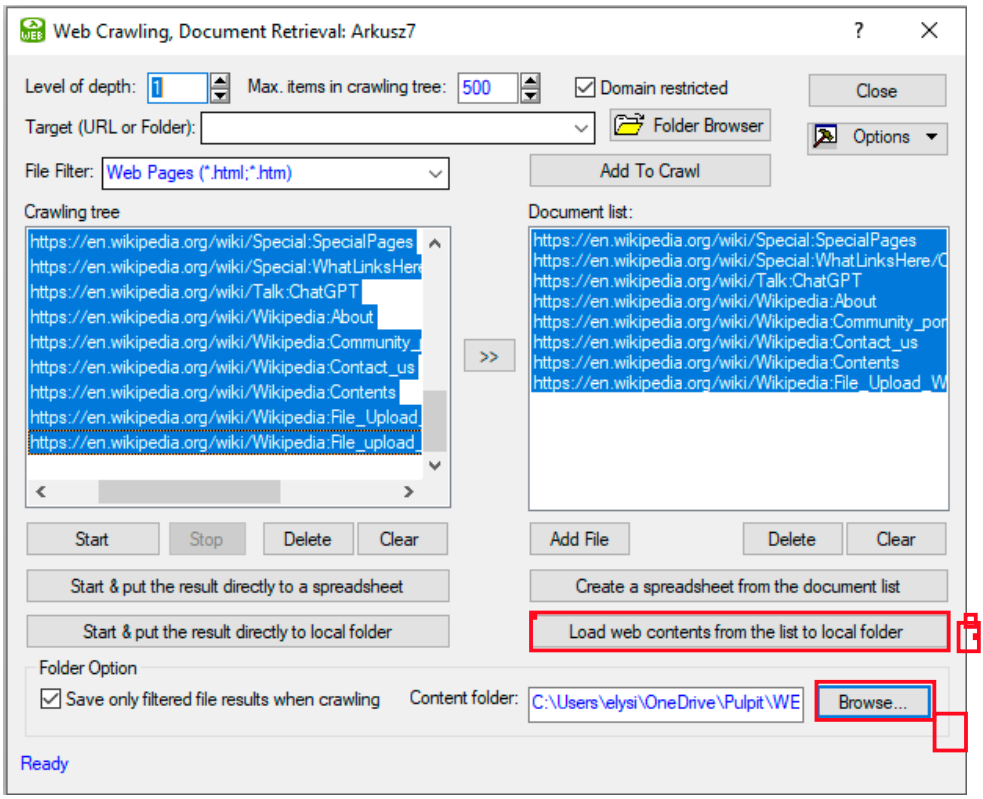


Figure 28.15. Web scraping procedure, steps 3 and 4

These pages can then be analysed by importing them into STATISTICA and performing various operations, such as text mining and other advanced analyses. With the capabilities offered by STATISTICA, users can thoroughly search and analyse the content of the downloaded pages, identifying key patterns and relationships in textual data.

For instance, using text mining techniques, one can extract valuable information such as the most frequently occurring words and phrases, as well as conduct sentiment analysis. In addition, the available tools allow for the creation of predictive models and clustering of data, which enables even deeper understanding of the analysed material.

This process also includes the possibility of processing visual and tabular data, further enriching the scope of analysis. Once the data has been imported into STATISTICA, users can take advantage of a wide range of algorithms and analytical methods that support decision-making based on the gathered information.

In this way, STATISTICA becomes a tool that enables comprehensive research and analysis, which is invaluable across many fields of science and business. As a result, even large volumes of data retrieved from the Internet can be effectively managed, analysed, and interpreted—opening up new research and application possibilities.

This chapter presented the procedure for performing web crawling and web scraping in STATISTICA, which allows for the automated acquisition and analysis of data within a single environment. This is a convenient solution, especially for researchers already working in STATISTICA, who wish to carry out the entire process—from data retrieval to analysis—without the need for external tools.

However, it is worth noting that alternative methods for performing this process exist, particularly using the Python programming language. These methods not only allow for the extraction of content from websites but also support interaction with dynamic web pages that require simulation of user behaviour.

If the researcher requires greater flexibility or wishes to access pages that involve login and dynamic queries, Python may be a more suitable tool. Nevertheless, for many applications—especially when data is to be analysed immediately within the STATISTICA environment—the approach presented in this chapter will be sufficient and convenient.

In summary, web crawling and web scraping can be carried out in various ways—the choice of tool depends on the research objectives, the nature of the data, and user preferences.

PART VI

Social Network Analysis and Graph Theory

CHAPTER 29

Introduction to Social Network Analysis

29.1. Definitions and Basic Concepts

Social network analysis, also referred to as network analytics, is an advanced methodological approach that focuses on identifying and interpreting relational structures among individuals, groups, and organizations. According to the description by Abramek and Rizun, this method is applied to the representation of social connections using graphs, which makes it possible to illustrate the characteristics of nodes and the relationships between them, thereby enabling a deeper understanding of cohesion and dynamics within network structures. Their words emphasize the significance of this technique: *"Social network analysis allows researchers to focus on the relationships among individuals, groups, or organizations. It uses drawings of social connection graphs to display the properties of nodes and the relationships between them. It is applied to study the structure of networks and the relationships within them, which determine their cohesion and dynamics. The assumptions of the SNA method are intended to describe and understand patterns of interaction occurring within networks and the influence of these patterns on individual members"*. (Abramek & Rizun, 2015)

In the context of network analytics, adjacency matrices play a particularly important role; they record the properties of nodes and the nature of the relationships between them, becoming the foundation for in-depth network analysis. Nodes, their attributes, and the relationships among them are systematically documented in matrices, allowing for detailed examination and interpretation.

Graphs, employed at the analytical level, serve to visualize and further analyse these interactions, thereby facilitating the effective development and interpretation of complex social networks. The use of graphs in SNA not only facilitates visualisation but also enables the quantification of relationships within social groups, leading to a deeper understanding of the dynamics and cohesion within the network (Borgatti et al., 2009).

In social network analysis, it is essential to understand that the focus lies in identifying nodes, which may be interpreted as individual or psychological characteristics of actors, analysing the attributes of these nodes, and investigating the relationships between them. The application of tools such as graph theory and matrix algebra enables the formalisation of these relationships, which supports detailed data analysis and facilitates the use of various statistical and graphical tools. The work of Borgatti and colleagues (2009) highlights the importance of structural network analyses in studying social dynamics and relationships (Borgatti et al., 2009).

Social network analysis enables the study of structures shaped both by strong, frequent interactions and by so-called "weak ties" (Granovetter, 1973), which play a significant role in the diffusion of information and innovation. Within research, the analysis focuses on three main aspects: a) identification of nodes and their characteristics, b) analysis of the attributes of these nodes, and c) examination of the relationships between them, which allows for a more comprehensive understanding of social structures and their impact on individual behaviour.

The elements of networks presented below are based on established literature in the field, including the work of Sierocki (2020).

Nodes

Nodes are the fundamental elements of social networks, representing entities such as individuals, groups, organisations, or other units capable of interaction. In the context of network analysis, nodes are the points through which information, resources, or influence flow. The characteristics of nodes, including their position and roles within the network, can determine their impact on the structure and functioning of the network, as well as on the modes of resource and information distribution. Nodes in social networks gain significance through the ties that define their position in the network structure and enable the analysis of social relational dynamics (Sierocki, 2020).

Attributes

Attributes are properties assigned to nodes or edges, which can be either quantitative or qualitative. They encompass a broad range of characteristics, such as demographic features (age, gender), professional aspects (position, skills), as well as psychological traits (personality, preferences). These attributes are essential to the analysis, as they may influence the structure of relationships and patterns of interaction within the network.

Relationships

Relationships are the links between nodes that define the structure of the network. These connections may take various forms, including friendships, professional associations, transactional ties, or communicative interactions. Relationships can also be classified as directed, where the direction matters (e.g., the flow of information from sender to receiver), or undirected, where the direction is irrelevant (e.g., mutual friendship). The intensity and type of these relationships often determine the dynamics of the network and can be analysed in terms of their impact on the flow of resources and information. As noted by Sierocki, the existence of a network is made possible by the relationships between individual nodes (Sierocki, 2020).

Networks

Social networks can be classified according to various criteria, such as structure, type of relationship, or dynamics of change. One approach is the distinction between open and closed networks. In social network analysis, networks are defined as sets of nodes and the relationships between them. They may take different forms, from simple homogeneous networks to complex multilayered structures, where different types of nodes are interconnected in various ways. Networks may be analysed in terms of their density, centrality, and other network metrics that provide insights into their structure and functionality.

Types of Networks

Networks vary depending on their structure and the objectives of the analysis. These may include open networks, which have no clearly defined boundaries and may expand dynamically, or closed networks with explicitly delineated borders, such as employee networks within a company. Different types of networks require different data collection methods and analytical techniques, which may affect the interpretation of research findings.

Network Boundaries

Network boundaries define which nodes and relationships are included in the analysis. The way these boundaries are defined can significantly influence the research results. Boundaries may be determined by the researcher (nominal) or emerge from the participants' natural awareness (realistic), which is of particular importance when defining the "whole network" in a study. One example is the distinction between a social network defined by participation in a specific social group and a network defined on the basis of professional contacts.

Structural Holes

Structural holes are a concept developed by Ronald Burt, describing situations in which there is a lack of connections between groups of nodes in a network (Burt, 2022). Individuals or groups occupying such gaps may act as intermediaries or information brokers, thereby connecting different, otherwise unlinked groups. Holding such a position within the network may provide strategic advantages, such as access to diverse resources and information not available to other network members. Structural holes introduce an additional dimension to social network analysis, offering insights into how the flow of information and resources may be controlled and how networks can be optimised to enhance innovation and efficiency.

Cliques and Clusters

As noted by Sierocki (2020), any network can be divided into a smaller number of categories such as cliques, components, or clusters, based on the characteristics of relationships between nodes rather than their attribute-based features. Cliques, in the context of social networks, are groups in which every node is directly connected to every other node within the group. This represents the most cohesive possible subset within a network, meaning that all potential links between the nodes in a clique actually exist. Cliques are often used to identify highly cohesive communities in networks, which may be relevant for the analysis of social or professional groups, where a high level of cohesion supports mutual assistance and collaboration.

Clusters, on the other hand, are similar to cliques in that they represent groups of nodes more strongly connected to each other than to the rest of the network, but they do not require full direct cohesion as cliques do. Clustering in networks enables the identification of natural groups, communities, or segments within larger networks, which is especially useful when analysing large data sets, where individual clusters may represent people with similar interests, professions, or other shared attributes.

Each of these concepts contributes uniquely to the ways in which researchers can interpret and manipulate network data to generate new insights and understand existing patterns on both the local and global levels.

29.2. History and Development of Social Network Analysis

Social network analysis (SNA) has undergone a long developmental journey from its early days to the present, evolving from simple interpersonal observations to advanced mathematical and statistical models. This chapter draws extensively on the work of Freeman, who describes in detail the history and development of social network analysis, including key concepts such as sociometry, sociograms, and the contributions of researchers such as Moreno and Lewin (Freeman, 2004).

In the 1930s, Jacob Moreno introduced the concept of sociometry, which constituted the first systematic approach to studying social networks using quantitative methods. This innovative approach enabled the identification of leaders, isolated individuals, and subgroups within communities, marking a breakthrough in understanding group dynamics (Freeman, 2004).

Sociometry, developed by Jacob Moreno, represented a methodological breakthrough in social research, relying on the quantification of interpersonal relationships and dynamic group structures. Moreno used this technique to measure attractions and rejections in interpersonal relationships, which allowed for the empirical analysis of interaction patterns. His sociometric method, described in detail in *Who Shall Survive?*, enabled researchers to identify key relationships and the roles of individuals within communities, which was fundamental to understanding how social structures influence individual behaviour.

Sociograms, introduced by Moreno as a tool for graphically representing social networks, revolutionised the way in which sociometric data were presented and analysed. These diagrams illustrate the relationships among members of a group, showing how individuals are connected through various types of ties. Each node in a sociogram represents a person, while the lines represent relationships, which may be further marked with arrows indicating the direction of the relationship (e.g., sympathy, antipathy). Thanks to sociograms, it became possible to quickly and visually read the social structure, identify leaders, isolated individuals, and observe how relationships evolve in response to interventions. This method not only facilitated a deeper understanding of group dynamics but also became a key tool in the development of subsequent network research (Freeman, 2004).

A sociogram is a graphical representation of social networks, composed of nodes and edges that illustrate the structure and dynamics of relationships among individuals. Each node in the sociogram symbolises a person or entity in the network, typically labelled with the person's name or another identifier. Edges, or the lines connecting the nodes, represent various types of relationships—these may be friendships, professional or family ties, or others defined by the researcher. The nature of these lines may vary; for example, dashed lines may symbolise weaker ties, whereas solid lines indicate stronger connections. Additionally, the directionality of the lines, represented by arrows, may indicate the direction of the relationship, which is crucial in studies of hierarchical or emotional relationships, where it is important to know who initiates the interaction. Groupings in a sociogram visualise how individuals placed close together and connected by multiple lines form subgroups or cliques within the larger network, which may reflect close or intense interactions.

Through the introduction of sociometry and sociograms, Moreno not only contributed to the development of research methods in social psychology and sociology, but also laid the foundation for social network analysis (SNA). These tools enabled researchers to precisely track and analyse complex relational patterns, which were later expanded into more advanced research techniques used in SNA. Moreno's work, by highlighting the structural and functional significance of relationships

within groups, opened new perspectives in the study of social networks—perspectives that continue to be explored today.

By applying these innovative methods, Moreno not only enhanced the understanding of how communities function but also initiated processes that became pivotal for the further development of social network analysis as an interdisciplinary and dynamically evolving scientific field.

The post-war period was a time when SNA had not yet been widely recognised as a distinct discipline; however, research on small groups—particularly the work of Kurt Lewin—contributed to a better understanding of group functioning mechanisms. Lewin studied the impact of various factors on group structures, which later found applications in network analysis.

Kurt Lewin, regarded as a pioneer of social psychology, introduced the concept of field theory, which had a significant influence on the later development of social network analysis (Lewin, 1952). Lewin's field theory proposed that an individual's behaviour results from dynamic interactions between the person and the surrounding "field of forces". This field consists of various social and psychological forces affecting the individual in a given context. Freeman emphasises how Lewin's field theory inspired researchers to explore how social structures influence individuals—a key component of social network analysis (Freeman, 2004).

Lewin also conducted pioneering studies on group dynamics, focusing on how group structures change under the influence of various internal and external factors. His experiments involving leadership styles revealed how changes in group structure affect decision-making processes and interactions among members. Freeman notes that Lewin's work on groups and leadership provided insights into how patterns of relationships shape group functioning, directly influencing the development of social network analysis methods.

Although Kurt Lewin did not directly employ methods of social network analysis, his approach to group research and his use of experimental and quantitative techniques laid the groundwork for later scholars to develop SNA. Freeman underscores how Lewinian research techniques influenced the formation of methods later adapted to analyse complex social systems within the SNA framework.

Kurt Lewin's scientific contributions—though not initially directly related to social network analysis—provided essential foundations that enabled the growth of the field. As Freeman observes, Lewin's concepts of forces within an individual's life space and his group research offered invaluable insights into mechanisms that would later become central to social network analysis.

The post-war period also marked a time when researchers such as Elizabeth Bott began exploring domestic relational networks, contributing to a deeper understanding of family interactions and their influence on broader social behaviours. Elizabeth Bott, known primarily for her research on family relationship networks published in her work *Family and Social Network* (1957), examined how domestic structures affect roles and dynamics between spouses. Her work shed light on the connections between network configurations and gender roles,

emphasising how the diversity and density of social networks influence the division of responsibilities and mutual support within families. Bott's research revealed that couples with broad, diverse relational networks often experience greater flexibility in negotiating domestic roles, which translates into more egalitarian solutions in daily life. These findings not only enhanced the understanding of the structure and function of family networks but also enriched the broader concept of social network analysis by demonstrating its relevance in practical, everyday social contexts (Freeman, 2004).

A breakthrough moment for SNA occurred in the 1960s, when Harrison White and his colleagues at Harvard began applying formal mathematical models to network analysis. They developed the concept of *blockmodeling*, which allowed for the classification of network nodes based on their structurally equivalent positions. This period also saw the first use of computers for network data analysis, significantly expanding research capabilities and leading to the development of more complex algorithms and analytical techniques (Freeman, 2004).

Harrison White, as one of the key figures in the development of social network analysis, contributed to the formalisation of the mathematical foundations of the discipline. His work during the 1960s and 1970s at Harvard University was pioneering in the use of mathematical models to study social structures. White and his collaborators—including many who later became leaders in the field—applied a range of innovative approaches that revolutionised SNA.

A central contribution of Harrison White was the introduction of the *blockmodeling* method, which enabled the analysis of social networks by grouping nodes based on the similarity of their connections. This technique allowed researchers to study large networks more effectively by reducing complexity and focusing on patterns of connections between various groups within the network.

Blockmodeling is based on the adjacency matrix, in which nodes are arranged so that those with similar patterns of connections are grouped together. In practice, this means that the matrix is reorganised in such a way that blocks (groups of nodes) with high internal density and low inter-group density become visible. This approach makes it possible to identify *positions* and *roles* within the network, which is essential for understanding the structural foundations of the network.

Another significant contribution by White was the application of algebraic methods to network analysis, which enabled a more formal and quantifiable approach to studying social structures. White's work employed, among other things, graph theory and various algorithms to analyse paths and accessibility within networks, leading to the development of new tools for measuring centrality, closeness, and betweenness in networks.

The work of Harrison White had a far-reaching impact on the development of social network analysis. His mathematical methods became the foundation for many subsequent studies and theories in SNA, including those of scholars such as Ronald Burt, Mark Granovetter, and Barry Wellman, who further developed and adapted these concepts in various social and economic contexts.

Harrison White's scientific contributions were groundbreaking, not only because of his introduction of new mathematical methods, but also due to his integration of those methods with the empirical investigation of real-world social networks. His work contributed to transforming social network analysis from a loosely defined concept into a fully-fledged, mathematically grounded scientific discipline, offering profound insights into the nature and dynamics of social relationships.

The 1970s and 1980s marked a period when SNA gained prominence as a formal scientific discipline. The first structured research programmes emerged, as well as tools such as UCINET, which enabled the processing of large data sets. During this period, SNA began to be applied not only in sociology but also in other fields such as anthropology, economics, and the emerging health sciences. Researchers such as Mark Granovetter introduced concepts like the "strength of weak ties", which revolutionised the understanding of how information and innovation spread through social networks (Freeman, 2004).

Social network analysis, from its modest beginnings in Moreno's studies to its broad contemporary applications, has become an essential tool for scholars seeking to understand complex social patterns and structures. The development of this field reflects its interdisciplinary nature and its ability to adapt to changing conditions and available technologies.

Today, social network analysis (SNA) is widely used across a variety of scientific and practical disciplines—from sociology and psychology to computer science, biology, epidemiology, and even marketing and management. The application of SNA has expanded due to the increasing availability of large data sets and advanced computational tools, which make it possible to analyse complex and extensive social networks.

In contemporary settings, SNA plays a key role in analysing patterns of connection and information flow in social media, allowing for the identification of influential users and the examination of content diffusion. In epidemiology, SNA is used to monitor and model the spread of diseases, while in knowledge management it facilitates the optimisation of information and innovation flows within organisations. SNA is also applied in biology, aiding the analysis of ecological networks, and in security sciences, where it supports law enforcement in mapping criminal and terrorist networks.

These versatile applications are made possible by the use of advanced mathematical and statistical tools. Graph theory, which forms the foundation of SNA, allows networks to be represented as sets of nodes and edges, which is essential for all forms of structural analysis.

In network analysis, adjacency and incidence matrices are also crucial, as they precisely describe how nodes are connected to one another. Graph traversal algorithms, such as breadth-first search (BFS) and depth-first search (DFS), are employed to examine accessibility and paths within networks. Measures of centrality are equally important, as they help identify the most influential nodes, while clustering techniques and community detection enable the identification of more tightly connected groups within a network.

SNA also employs advanced statistical models that allow for the examination and interpretation of complex relational patterns in networks. These models—including latent space models and mixed-effects models—make it possible to rigorously test hypotheses concerning the structure and dynamics of networks. The introduction of these mathematical methods into network research, initiated by pioneers such as Harrison White, revolutionised SNA by enabling a shift from intuitive research methods to reliable, quantifiable analyses.

Thanks to these tools, SNA has become not only a research method used in sociology and psychology but one that has found application across many other fields, reflecting its interdisciplinary nature and adaptability. The development of this discipline demonstrates how analytical methods can evolve in response to new challenges and technologies, providing researchers with powerful tools for exploring complex social systems.

CHAPTER 30

Basic Methods and Tools in Network Analysis

30.1. Graphs and Their Representations

Social network analysis (SNA) extensively utilises graph theory as a foundation for modeling social relationships and structures. In the academic literature, graphs constitute a universal tool for visualising and analysing complex social networks. This chapter presents the basic types of graphs used in SNA and their applications, based on the general principles of graph theory and scholarly sources highlighting their significance in research and practical contexts. The information presented here draws on widely accessible literature and application examples found in the works of Alamsyah, Rahardjo, and Kuspriyanto (2013), Umami, Prihandini, and Agatha (2024), and Szymańska (2024b). Below, selected types of graphs, their characteristics, and examples of their applications in social network analysis are described.

Social network analysis (SNA) is based on the use of graphs as fundamental tools for modeling complex relational structures. The graphical representation of networks enables researchers to intuitively understand social structures, group dynamics, and mechanisms of social influence, which are essential in various research and application contexts (Alamsyah et al., 2013). Depending on their characteristics and the requirements of the analysis, graphs may take different forms, allowing for an appropriate representation of network dynamics and structure. In social network analysis, the most commonly used graphs are: a) undirected, b) directed, c) weighted, as well as special types of graphs such as d) bipartite and e) hypergraphs.

In *undirected graphs*, the relationships between nodes are reciprocal, meaning that the edges have no direction (Umami et al., 2024). These types of graphs are

used to model networks in which the relationships are symmetrical, such as friendship networks. In undirected graphs, relationships between nodes are mutual, which means the direction of interaction is irrelevant. They are particularly useful for modeling ties such as friendships or collaborations, where the interaction is bidirectional. Examples of undirected graphs in social research can be found in the literature, where their versatility in network analysis is emphasised—for instance, in studies on relationships between personality disorders (Szymańska, 2024c).

In *directed graphs*, edges have a specific direction, meaning the relationships between nodes are not reciprocal. This type of graph is used to model networks in which the relationships are directional, such as influence networks (Umami et al., 2024). Directed graphs, also known as *digraphs*, have edges with a defined orientation, which is crucial in modeling structures where the direction of flow matters, such as in organisational hierarchies or information networks.

In *weighted* graphs, edges are assigned weights that represent the strength or intensity of the relationships between nodes. This type of graph is used in analyses where the differentiation in relationship strength is essential, for example in communication networks. In weighted graphs, the edges carry weights that illustrate the strength or intensity of a relationship—such as the frequency of interaction or the volume of transactions. These graphs enable a deeper analysis of the significance of specific connections within the network (Umami et al., 2024).

Bipartite graphs consist of two distinct sets of nodes, where edges connect only nodes from different sets (Figure 30.1). These types of graphs are used to model relationships between two different types of entities, such as people and events (Umami et al., 2024).

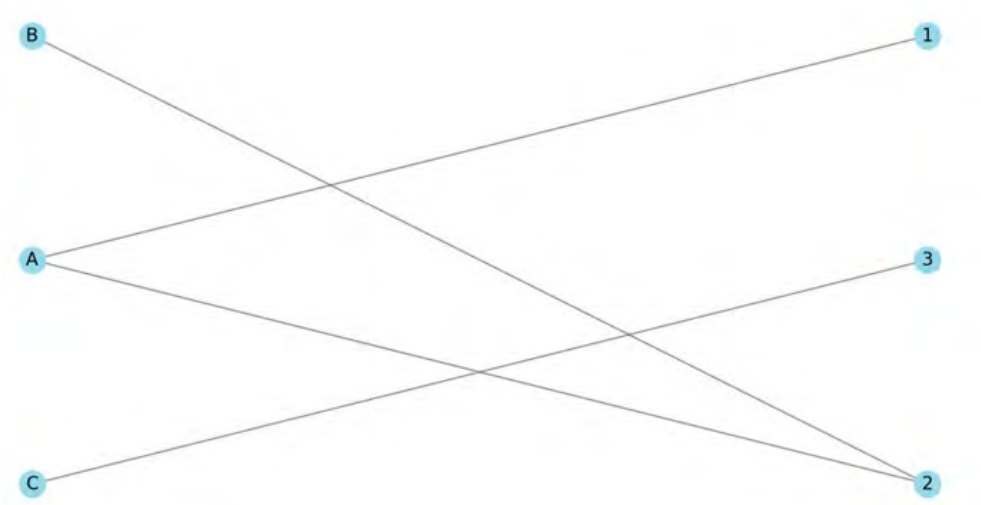


Figure 30.1 Example of a Bipartite Graph

In hypergraphs, edges may connect more than two nodes (Figure 30.2). This type of graph is used in more complex analyses where relationships may involve groups of entities (Alamsyah et al., 2013). Hypergraphs allow for the representation of relationships in which edges can simultaneously connect more than two nodes. This is useful in situations where group relations cannot be adequately represented by simple one-to-one connections.

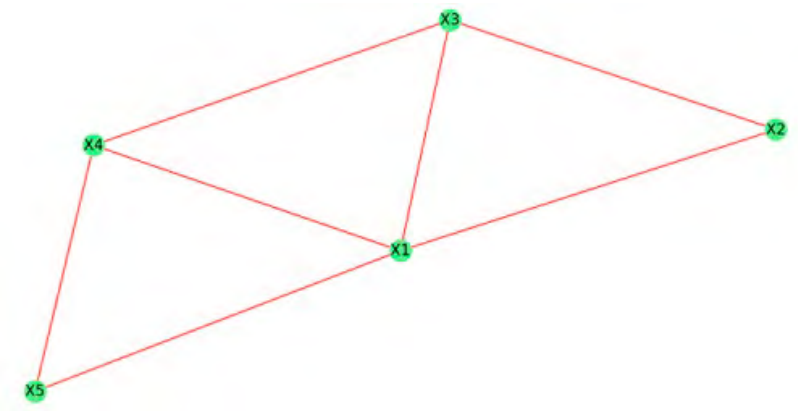


Figure 30.2 Example of a Hypergraph

Graph Representation Methods

Graph representation in SNA is essential for the visualisation and analysis of data. Several commonly used graph representation methods support the analysis of network structure and dynamics.

The adjacency matrix is one of the basic forms of graph representation, where rows and columns represent nodes, and the values in the matrix indicate the presence (and possibly the weight) of edges between nodes (Chakraborty et al., 2018). An adjacency matrix is a square matrix used to represent graphs, with rows and columns corresponding to nodes and the cell values indicating whether an edge exists between them. It is a fundamental form of graph representation, particularly useful in algebraic analysis.

An alternative to the adjacency matrix is the adjacency list, which stores for each node a list of other nodes to which it is directly connected (Chakraborty et al., 2018). This method is more efficient in the case of sparse graphs, as it consumes less memory than the adjacency matrix. The adjacency list represents a graph using lists, where each node is associated with a list of nodes it connects to. This form of representation is memory-efficient and commonly used in the analysis of large networks.

Graphical representation is invaluable in visualising network structures, facilitating the identification of patterns, key nodes, or isolated segments. The academic literature emphasises the importance of visualisation in network analysis, providing

examples of how graphical representations of graphs can support the analysis and interpretation of complex network data (Umami et al., 2024).

For the analysis of networks that change over time, dynamic representations are used, which allow researchers to track the evolution of a graph. Dynamic graphs are particularly important, as many social networks are not static and evolve over time. Dynamic representations enable researchers to observe how network structures change, how connections between nodes shift, and how key nodes emerge and disappear. Through dynamic graphs, it is possible to model and analyse processes such as the diffusion of information, changes in organisational structures, or dynamic interactions in social networks (Butts et al., 2024).

Algorithms Used in Graph Analysis in SNA

Social network analysis (SNA) employs a variety of algorithms that enable the investigation of complex relational structures and network dynamics. Among these algorithms, several key ones are widely used for different analytical purposes.

Information flow analysis algorithms are used to study how information moves through a network, which is essential in the analysis of social, communication, and biological networks. The most important algorithms include the PageRank algorithm and the HITS (Hyperlink-Induced Topic Search) algorithm. PageRank evaluates the importance of nodes based on the structure of connections, whereby nodes with a high PageRank are considered significant within the context of the entire network (Stoica et al., 2024).

It is worth noting, however, that these algorithms may reflect and reinforce existing biases in networks. PageRank, while reflecting the degree distribution within the network, can balance the representation of minority groups among the highest-ranked nodes. In contrast, HITS, especially in homophilic networks, tends to amplify existing biases—an effect that has been demonstrated both theoretically and empirically (Stoica et al., 2024).

Although these algorithms are highly useful for network analysis, they may not always operate in a neutral manner. PageRank functions such that the importance of a node is proportional to the number and quality of its links to other nodes. Consequently, in diverse networks, this algorithm may enhance the representation of minority groups among the most central nodes. This is particularly relevant in the context of social networks, where access to information or influence should not be dominated solely by the majority.

On the other hand, HITS operates differently by distinguishing between two types of nodes: authorities and hubs. In networks characterised by homophily (i.e., nodes are more likely to connect with similar others), this algorithm reinforces existing inequalities. This means that groups already well connected become even more prominent, whereas smaller or less connected groups lose visibility. Stoica and colleagues (2024) have shown that this phenomenon may be especially problematic in highly segregated networks, as it leads to further marginalisation of minority groups.

As a result, the choice of algorithm should depend on the context and the goal of the analysis. If balancing the representation of different groups within the network is important, PageRank may be the more suitable choice. In contrast, in situations where a hierarchical structure of authorities and hubs is central, HITS may yield more relevant results—though its limitations must be acknowledged.

Another group consists of **clustering and structural analysis algorithms**, which are used to identify patterns and structural properties within a network. These algorithms differ from methods such as PageRank or HITS, which focus on analysing the flow of information and the importance of individual nodes. Clustering focuses on grouping nodes based on their similarities or the intensity of their connections within the network (Khan & Niazi, 2017). Its goal is to identify *communities*, i.e., groups of nodes that are strongly connected internally but have few links to other parts of the network.

In contrast to information flow algorithms, which analyse the direction and significance of resource exchange between nodes, clustering algorithms examine the overall structure of the network. For example, PageRank assesses a node's importance based on the quality and number of its links, while clustering algorithms—such as *k-means* or *Louvain*—aim to understand which nodes naturally form groups. Thus, clustering methods make it possible to detect communities or patterns that may be invisible in traditional flow-based analyses.

The simplest clustering algorithms, such as *k-means*, operate on the basis of similarity measures between nodes. K-means assigns nodes to clusters based on their attributes and iteratively optimises these groups to minimise intra-cluster differences and maximise inter-cluster differences. However, *k-means* does not account for graph structure, which makes it less useful in the context of complex social networks. To more effectively analyse such networks, more advanced methods are applied, such as *spectral clustering*, which uses eigenvalues of the graph's adjacency matrix (Khan & Niazi, 2017).

Community detection refers to the process of identifying groups of nodes that are more densely connected to each other than to the rest of the network. Common algorithms in this area include the *Louvain* algorithm and the *Girvan–Newman* algorithm. The Louvain algorithm is one of the most popular community detection methods; it maximises network modularity by dividing the network into communities (Khan & Niazi, 2017). The Girvan–Newman algorithm, by contrast, identifies communities by progressively removing edges with the highest betweenness centrality, which results in the network being split into smaller groups (Khan & Niazi, 2017).

Clustering and structural analysis algorithms provide valuable tools for examining social networks, allowing for a deeper understanding of their internal structure. Their use enables the extraction of communities, the analysis of patterns, and the identification of key groups within a network. The diversity of methods—from simple algorithms like *k-means* to advanced spectral techniques—makes it possible to tailor analytical tools to the specific nature of the network under study and the goals of the analysis.

In summary, the algorithms used in graph analysis within SNA are crucial for understanding the structure and dynamics of social networks. They make it possible to identify key nodes, detect communities, examine the flow of information, and analyse the structural properties of networks. The use of appropriate algorithms allows researchers to gain deeper insight into complex social relations and to model networks effectively.

Practical Applications of Graphs in SNA

Graphs are employed in numerous practical applications of social network analysis (SNA). One of the main applications is community detection, i.e., identifying groups of nodes that are more densely connected with each other than with the rest of the network (Alamsyah et al., 2013). This type of analysis enables researchers to understand how groups of individuals or entities interact with one another, which may be useful in studies on group dynamics and social structures.

Another important application is centrality analysis. Centrality measures the importance of individual nodes in a network based on various metrics, such as degree centrality, betweenness centrality, and closeness centrality (Arif, 2015). These metrics help identify key nodes that may play a significant role in information flow, social influence, or the coordination of activities within the network.

Information flow analysis constitutes another essential aspect of applying graphs in SNA. Investigating how information circulates through a network can be relevant in the analysis of social, communication, and biological networks. This allows researchers to understand how information, resources, or diseases spread across networks, which is crucial in various fields, ranging from public health to marketing and management.

Graph representations and various methods of their analysis are indispensable in understanding and modeling complex social networks. They enable researchers to effectively analyse and interpret data, which is essential for drawing conclusions and formulating data-driven strategies.

30.2. Metrics in Social Networks

In social network analysis (SNA), network metrics play a crucial role in understanding the structure and dynamics of networks. These metrics allow for the quantitative description of the position and importance of individual nodes as well as entire networks. Among the most important metrics, we distinguish: a) centrality, b) cohesion, c) information flow, and d) community detection.

Centrality is one of the most essential concepts in social network analysis. It evaluates the significance of individual nodes in the network, enabling the identification of key actors. Several types of centrality are distinguished: degree centrality, betweenness centrality, and closeness centrality (Arif, 2015). Degree centrality is the

simplest measure, which counts the number of direct connections (edges) of a given node. Nodes with high degree centrality are well-connected and may play key roles in information distribution (Arif, 2015). Betweenness centrality measures how often a given node lies on the shortest path between two other nodes. Nodes with high betweenness centrality are essential for information flow in the network, acting as intermediaries (Arif, 2015). Closeness centrality measures how close a given node is to all other nodes in the network. Nodes with high closeness centrality can disseminate information quickly within the network (Arif, 2015).

Network cohesion is another important metric in SNA. It defines how well the nodes in a network are connected. Network density refers to the ratio of existing edges to the maximum possible number of edges in the network, indicating the intensity of connections within the network (Khan & Niazi, 2017). The clustering coefficient indicates the extent to which nodes in a network tend to form tightly connected groups (clusters). A high clustering coefficient suggests the presence of strongly interconnected groups in the network (Khan & Niazi, 2017). Average path length measures the average length of the shortest path between pairs of nodes, with shorter paths indicating a more cohesive network.

Information flow concerns the manner in which data or influence spreads within the network. Metrics such as global efficiency and local efficiency allow for the assessment of how effectively information reaches nodes, taking into account the structure of the entire network.

Community detection involves identifying subgroups in a network whose members are more strongly connected to each other than to the rest of the network. These communities may correspond to actual social groups, organizational teams, or other functional units.

These metrics, when applied together with appropriate algorithms, enable a precise understanding of the structure and dynamics of social networks, making it possible to identify key nodes, information flows, and social groups.

CHAPTER 31

Practical Applications of Social Network Analysis

31.1. Network Analysis in the Social Sciences

31.1.1. Social Network Analysis in Psychotherapy and Upbringing Psychology

This chapter describes the application of social network analysis (SNA) in clinical psychology, psychotherapy, and upbringing psychology. Social network analysis enables the understanding of complex patterns of interpersonal interactions, which is essential for designing effective psychotherapeutic interventions.

SNA offers possibilities for examining dynamics within therapeutic groups by analysing how changes in network structure can affect the effectiveness of psychotherapy. For example, it allows the analysis of how the formation and dissolution of support groups influence the outcomes of psychotherapeutic interventions, making it possible to apply more targeted and effective approaches in group psychotherapy. As demonstrated in the analysis conducted by Szymańska, it enables the identification of relationships between personality disorders (Szymańska, 2024c).

The development and application of SNA methods in psychology allow for a deeper understanding not only of the structure and dynamics of social networks but also for the development of more effective intervention methods that are tailored to the specific dynamics within the patients' social networks.

In a study conducted by Szymańska and Aranowska, the analysis of parental goals and the development of personality traits in preschool children was carried

out using advanced methodologies, including Social Network Analysis (SNA). The study focused on the traits that parents aim to foster and those they actively attempt to inhibit in their children, using a comprehensive framework that also included text exploration, support vector machines, and the Aranowska's compatibility coefficient (λ) for data analysis (Szymańska & Aranowska, 2022).

The results revealed that parents prioritise the development of competence-related traits, particularly independence, and attempt to prevent the development of negative temperament traits, such as aggression. These traits and parental goals change with the child's age, indicating a dynamic approach to parenting that is influenced by the child's developmental stage.

This nuanced approach allows for an in-depth understanding of the complex interactions between parental aspirations and the social and temperamental development of children, providing valuable insights into how parents can effectively shape their child's personality through conscious goal-setting and the avoidance of undesirable traits.

31.1.2. Studying Social Phenomena through Network Analysis: From Mental Health to Organizational Dynamics

Social Network Analysis (SNA) is an interdisciplinary tool that enables the understanding of complex dependencies and dynamics within social structures by employing advanced graph-based and mathematical methods. SNA is widely applied in the social sciences to study a range of phenomena, such as mental health, risk behaviours, workplace dynamics, and personality development in the context of social networks. Numerous examples from the literature demonstrate how SNA has been utilised to explore various issues in psychology.

One example of applying Social Network Analysis in psychology is the study by Luo et al. (2022), in which the authors used stochastic actor-oriented models (SAOM) to analyse how evolving communication networks in interdisciplinary scientific teams influence participants' perceptions of psychological safety. This study shows how social interactions within dynamically developing networks can impact individual psychological states, such as the sense of safety in team relationships (Luo et al., 2022).

A systematic review conducted by Jeon and Goodson (2015), based on the analysis of 15 studies using AddHealth data and Social Network Analysis (SNA) methods, revealed that friendship networks play a significant role in shaping adolescents' risky behaviours, such as alcohol consumption, smoking, sexual activity, and marijuana use. The review findings contributed to a better understanding of how peer interactions influence adolescents' health-related decisions (Jeon & Goodson, 2015).

Further studies demonstrated how centrality in social networks affects the recovery processes of individuals with mental disorders, showing that individuals more embedded in the relational structure—those with more connections and occupying

more central positions—more frequently reported improvements in mental health. This underscores the importance of the social environment and relational structures in the context of psychological support (Ma & Sayama, 2015).

In studies on rhesus macaques, social network analysis provided insights into how social status affects health, indicating the applicability of SNA beyond human contexts (Vandeleest et al., 2016). Research on the influence of friendship networks on students' engagement in academic behaviours highlighted the role of peers in education (Shin, 2022).

Additionally, social network analysis (SNA) has been used to model the relationships between individual network positions and the day-to-day variability of personality states, allowing for a better grasp of individual functioning dynamics in the context of social relationships (Lekkas et al., 2022). Studies on children's early social environments, employing SNA to analyse the formation of early social experiences, provided valuable insights into children's behaviour and cognitive processes (Burke et al., 2022).

Further studies revealed how networked relationships in the workplace influence innovation and productivity, emphasising the importance of social structures in organisational contexts (Letouche & Wille, 2022). These examples from the literature illustrate the diversity of SNA applications in the social sciences, demonstrating its potential for investigating social and behavioural phenomena across various contexts and environments.

CHAPTER 32

Advanced Methods in Network Analysis

32.1. Community Detection in Networks

Community detection in networks, a key area of research in network analysis, focuses on identifying groups of nodes that exhibit stronger interconnections with one another than with the rest of the network. It is an indispensable method for understanding the complex structures and dynamics present in social networks, both in scientific and practical contexts, and finds applications in various fields such as sociology, bioinformatics, and knowledge management.

In practice, community detection helps to understand how elements within a network form subgroups that are more cohesive internally than in their relations with the rest of the network. For instance, in social networks, these groups may correspond to communities of acquaintances who communicate intensively among themselves while having less frequent contact with other groups. Similarly, in organisations, it is possible to identify teams that collaborate more closely within their units than with other departments.

This method relies on analysing the links between nodes (e.g., individuals, genes, devices) and identifying those that form densely interwoven networks. It is crucial for understanding how groups function within various systems. For example, in research on social relations, this method can help uncover which support groups form within local communities or which subcultures emerge in larger populations.

Various algorithms are used in community detection to automate this process by analysing network structures. Algorithms such as Louvain or spectral techniques examine data to identify groups with the greatest internal cohesion. Some

of them, such as hierarchical algorithms, allow for the division of the network at different levels of granularity, thus enabling the analysis of both smaller subgroups and larger structures.

Community detection has broad applications in psychology, for instance in analysing the dynamics of therapeutic groups. It may aid in identifying natural group leaders, support relationships among participants, or subgroups that may require special attention. In schools, this method can be used to analyse relationships among students, helping to identify peer groups and understand how they influence student behaviour.

In summary, community detection is not only an analytical method but also a tool that allows for a deeper understanding of the mechanisms governing group interactions. Its interdisciplinary nature makes it applicable across many fields, from the social sciences and biology to data analysis in enterprises. In psychology, it may serve as a valuable support in studying interpersonal relationships and group dynamics.

32.2. Multilayer and Complex Networks

Multilayer networks constitute an advanced tool in the analysis of complex social, biological, and technological systems. They are characterised by the inclusion of multiple layers, each representing different types of relationships between the same nodes. In such networks, each layer may have its own unique nodes and edges, and interlayer edges may also exist, linking nodes across different layers (Meštrović et al., 2022).

Multilayer networks can be categorised into different types, including multiplex networks, in which various layers represent different types of interactions between the same nodes. For example, one layer may correspond to professional relationships, while another to social interactions. In multiplex networks, nodes remain constant across all layers, which allows for the simultaneous modeling of multiple aspects of interactions between individuals in complex systems, such as social networks, where the same people coexist in different relational contexts (Boccaletti et al., 2014).

In a multiplex network, a node representing an antisocial personality disorder can be analysed across different layers, each corresponding to a specific aspect of functioning. In one layer, the relationships between this disorder and other mental disorders—such as sadistic, narcissistic, or histrionic personality—can be examined. In another layer, its impact on the individual's performance in social roles—such as professional relationships or participation in social life—can be analysed. Yet another layer may focus on family relations, such as interactions with a partner, children, or parents.

Various methods are used to analyse multilayer networks, including supra-adjacency matrices, which integrate all layers into a single matrix, thereby facilitating the

analysis of the entire network (Kivelä et al., 2014). Other techniques, such as "flattening", allow for the simplification of a multilayer network into a single-layer one while preserving key structural information (Kivelä et al., 2014).

Examples of multilayer network applications span various fields. In social research, such networks enable the analysis of different types of relationships, such as professional, social, or communicative ties, making it possible to study their mutual influences. In biology, multilayer networks can represent interactions between different types of molecules, e.g., proteins and genes. In transportation systems, different layers may represent various modes of transport, such as roads, railways, or air travel networks (cf. Kivelä et al., 2014). They also find applications in biomedicine, ecology, and economics (Boccaletti et al., 2014).

Multilayer networks offer numerous advantages over traditional single-layer networks. Most notably, they allow for a richer representation of data, enabling the capture of diverse types of relationships, which leads to a more detailed depiction of interactions. Analysing the layers either separately or collectively makes it possible to study the influence of different types of relationships on the network's structure and dynamics. The ability to examine multiple layers facilitates the identification of dependencies and correlations between various types of interactions.

In summary, multilayer networks constitute an advanced tool for analysing complex relational systems. They enable the modeling and analysis of various types of interactions within a single coherent framework, leading to a better understanding of complex social, biological, and technological systems. Thanks to advanced software tools, it is possible to efficiently analyse and visualise these networks, opening new research and practical opportunities.

Complex networks are structures found in various social, biological, and technological systems. They are characterised by intricate topological and dynamic properties that are absent in simpler networks, such as regular or random networks. The analysis and modeling of such networks allow for better understanding and prediction of the behaviours of systems in which these networks occur.

Complex networks consist of nodes and edges connecting them, forming a sophisticated structure. A key aspect is that such networks often exhibit features such as the "small-world" property and scale-freeness. Small-world networks are characterised by short paths between any two nodes and a high clustering coefficient, meaning that nodes tend to form densely connected groups (Svenson, 2006). Scale-free networks, on the other hand, follow a power-law degree distribution, which means that a small number of nodes have a very large number of connections, while the majority of nodes have relatively few (Svenson, 2006).

In practice, complex networks have applications across many domains. For example, in the analysis of social media such as Twitter or Facebook, these networks help identify influential users and study how information spreads through the network. This makes it possible to plan marketing campaigns and communication strategies more effectively (Malik, 2022). In biology, complex networks are used to model interactions between genes, proteins, and other molecules. Analysing such

networks can lead to the discovery of key elements in biological processes, which is crucial in medical and biotechnological research. In technological systems, such as computer or energy networks, complex networks assist in analysing system stability and resilience against failures and attacks. Through the modeling and analysis of complex networks, more reliable and efficient systems can be designed.

Various methods and tools are used to analyse complex networks. In centrality analysis, which evaluates the importance of individual nodes in a network, metrics such as degree centrality, betweenness centrality, and closeness centrality are employed (Malik, 2022). To recall, degree centrality counts the number of direct connections of a node, betweenness centrality measures how often a given node lies on the shortest path between two other nodes, and closeness centrality assesses how close a node is to all other nodes in the network.

Examples of complex networks can be found across many fields. In social media, such as Facebook or Twitter, the nodes are users and the edges represent their interactions. In biology, gene or protein interaction networks are crucial for understanding biological processes. In technology, computer networks—where the nodes are devices and the edges are the connections between them—are essential for ensuring the stability and security of the system. In transportation systems, airline or railway networks—where the nodes are airports or stations and the edges are the routes—enable efficient transport management.

Complex networks offer advanced possibilities for the analysis and modeling of complex social, biological, and technological systems. By applying appropriate methods and tools, it is possible to gain a precise understanding of the structure and dynamics of these networks, leading to improved management and optimisation of the systems under study.

CHAPTER 33

Software and Tools for Network Analysis

33.1. Overview of Network Analysis Tools

In social network analysis (SNA), tools and software play a key role in the visualisation and analysis of network data. They enable the efficient processing, analysis, and interpretation of complex relational structures. The development of tools for social network analysis has evolved over decades, beginning with simple manual methods and progressing to the advanced computer programs used today. The key stages of this development can be divided into several important steps.

In 1983, Franz Urban Pappi and Peter Kappelhoff from the University of Kiel created the SONIS program, and Linton C. Freeman at the University of California, Irvine, developed the first version of the UCINET program, which was later refined by Bruce McEvoy, Stephen P. Borgatti, and Martin G. Everett (Freeman, 2004).

In the 1980s, programs such as STRUCTURE and GRADAP—developed by Ronald Burt and by Robert Mokken, Felix Stokman, and James Anthonisse—were modified to include a broader range of analytical procedures. These were among the first attempts to create general-purpose programs for social network analysis that could be used by researchers across various disciplines (Freeman, 2004).

In the 1990s, the development of network analysis tools accelerated with the emergence of new computer technologies. Programs like UCINET became more accessible and widely used within the research community, which significantly influenced the development of the field. During this time, new programs such as PAJEK also emerged, which were specifically designed for analysing very large networks (Freeman, 2004).

Today, tools for social network analysis are significantly more advanced and offer a wide range of analytical and visualisation functions. Among the most commonly used tools and software in SNA are GEPHI, PAJEK, and CYTOSCAPE. GEPHI is one of the most popular tools for social network analysis. It is open-source software that enables the visualisation and analysis of large networks. GEPHI offers various functions such as filtering, clustering, centrality analysis, and dynamic network visualisation. Thanks to its intuitive user interface, GEPHI is widely used by both researchers and practitioners for exploring and analysing social networks (Apostolato, 2013).

PAJEK is another popular tool used in social network analysis. It is software specialised in analysing very large networks (Apostolato, 2013). PAJEK offers advanced analytical functions such as community detection, centrality analysis, hierarchy visualisation, and dynamic network analysis. PAJEK is especially valued for its efficiency and ability to process enormous volumes of data, which makes it an ideal tool for analysing large and complex networks.

NetDraw is software developed by Steve Borgatti in 2002, designed for visualising both one-mode and two-mode network data. It supports various file formats, such as UCINET, Pajek, and its own VNA format, which enables the storage of both network data and node attributes—such as their colours or sizes. The VNA format also allows the use of textual instead of numerical data, simplifying the network description. NetDraw supports multiple simultaneous relations and enables social network analysis, customisation of node appearance, and the export of graphics in formats such as JPG, GIF, or BMP. Thanks to its built-in functions, NetDraw partially overlaps with UCINET's analytical capabilities, while offering more extensive options for visualisation and data organisation. Detailed documentation and technical support are available on the program's website (Apostolato, 2013).

Each of these tools offers unique features and capabilities that allow for effective analysis and visualisation of social networks. They are indispensable tools in research on networks, enabling researchers to understand the complex structures and dynamics of social networks. These tools make it possible to conduct advanced analyses, identify key nodes, detect communities, and examine the flow of information within networks. Their advanced functions and capabilities allow researchers to efficiently process and interpret complex relational structures, which is essential for understanding and modeling social networks.

Selected tools for multilayer network analysis, such as Pymnet, Multinet, multiNetX, and muxViz, offer diverse modeling and analytical capabilities. These tools have been thoroughly compared in terms of performance and functionality in the context of multilayer network modeling (Ocklind, 2023).

33.2. Practical Workshops and Tutorial

This chapter, devoted to practical workshops and tutorials, focuses on presenting three diverse examples of the application of social network analysis (SNA) in the

study of various aspects of clinical and upbringing psychology. Each example employs a different methodology and analytical tool, allowing for an in-depth understanding of the potential of this technique.

The first example centres on the analysis of psychological traits that parents wish to develop in their children, both girls and boys. The UCINET program was used to illustrate how specific traits, treated as network nodes, are preferred depending on the child's gender, and how these traits are perceived by mothers and fathers.

The second example explored upbringing dynamics through network-based models of personality traits shaped by parents at different stages of their children's development. Numerical data, transformed into values corresponding to metatraits of personality, enabled the analysis of changes in parental preferences as the children matured. As in the first example, the classical UCINET program was used for data processing and network visualisation.

The final example demonstrates the use of advanced OpenAI artificial intelligence algorithms within the GPT language model, which autonomously synthesised and analysed complex textual data. This process allowed for the construction of an adjacency matrix and its corresponding SNA graph, illustrating the intricate connections and shared features among various personality disorders according to the DSM-IV classification (Millon & Davis, 1996).

Each of these examples shows how network analysis can be used to understand complex patterns and dynamics in social and clinical psychology, and how different tools and methods may be applied depending on the specific nature of the problem under investigation.

33.2.1. First Example: Selection of parenting goals, that is, psychological traits that parents actively seek to develop in their daughters and sons

In this chapter, we focus on the application of Social Network Analysis (SNA) in psychology. SNA is a powerful tool that enables the analysis of relationships and interactions between individuals within a network, allowing for a deeper understanding of the structure and dynamics of social groups. SNA methods facilitate the identification of key actors in the network, the determination of node centrality, and the analysis of communication and influence patterns.

To illustrate the practical application of SNA, we will refer to the examples and data discussed in Section 21.8.1, where parental goals adopted by mothers toward their children were analysed. These data, processed using text mining techniques, were used to generate matrices that enabled word counting and co-occurrence identification. This made it possible to determine which psychological traits are central in the structure of parental goals, as well as how mothers group different traits in the upbringing process.

Social Network Analysis in the upbringing context provided deeper insight into how mothers perceive their children's psychological development and which values and attitudes are most important to them. It also enabled a comparison of differences in parental goals with respect to sons and daughters, and the identification of possible gender-related patterns in upbringing approaches.

The SNA methods and results presented in this chapter offer valuable insights that may be utilised both in psychological practice and in further research on parenting and child development. In this context, a psychological trait constitutes a node in the analysed network. This could be, for instance, wisdom, patience, openness, etc. Node attributes include the frequency of a given trait in the population, reflecting how often it is mentioned by parents. Another attribute may be the child's gender for whom the trait is preferred (boy or girl). Other attributes may indicate whether the trait is more frequently mentioned by mothers or fathers, and whether it is considered desirable or undesirable.

A node representing a psychological trait that parents wish to cultivate may possess multiple attributes. One of the key attributes in network analysis is frequency of occurrence, which often correlates with popularity. This allows for the identification of central nodes in the network, such as those associated with high popularity or authority. Relationships between nodes may include various forms of membership in common sets. One example is the classification of traits as desirable or undesirable by parents. These relationships may also be defined by the order in which traits are indicated (e.g., listed first), as well as by membership in the category of desirable or undesirable traits.

Desirable and undesirable traits may form separate networks. For instance, traits mentioned in the first position may form one network, while those mentioned in the second position may form another. In this way, we can identify different network structures and understand how various traits are interconnected. In summary, the node, attribute, and relation are the three key elements contained in matrices that serve as the foundation for building graphs in network analysis. The information regarding attributes and relations may be either qualitative or quantitative in nature.

Let us examine a sample matrix. Table 21.1 (Section 21.8.1) presents a matrix in which the *nodes* represent parental goals adopted for girls, as well as the *attributes* of those nodes and the *relations* among them. The node "independence" was the most frequently mentioned first desirable trait that parents wished to cultivate in girls. Its frequency of occurrence reached as many as 45 instances. Between "independence" and "kindness", which constitutes the second node, there is a significant gap in popularity. "Kindness", as the second node, was mentioned 9 times as the most desirable trait. The third node is "self-confidence", mentioned 8 times, followed by "courage" – 7 times, "empathy" – 6 times, and "autonomy", "happiness", "values", and "faith" – each 5 times. Relations among these nodes may involve various aspects, such as their popularity hierarchy or classification as desirable or undesirable traits. This approach enables an in-depth analysis of social networks and the identification of key traits and their mutual connections in the context of parental preferences.

Table 21.2 (Section 21.8.1) presents a matrix for the nodes representing parental goals adopted by mothers for boys, as well as the attributes of those nodes and the relations among them. In the group of parents raising boys, independence was likewise the most typical trait, mentioned in the first position as many as 44 times. This was followed by empathy (10 times), self-confidence (8 times), creativity (8 times), courage, obedience, and faith (6 times each), and assertiveness, intelligence, and honesty (5 times each).

The table provides information on the node (the name of the cultivated trait) and four of its attributes: (a) type of goal – desirable or undesirable, (b) order of the listed trait, (c) frequency, and (d) child's gender. These data enabled the construction of a very simple network. Figure 33.1 presents the network for goals adopted toward girls, and Figure 33.2 for those adopted toward boys. The UCINET program was used to perform this social network analysis.

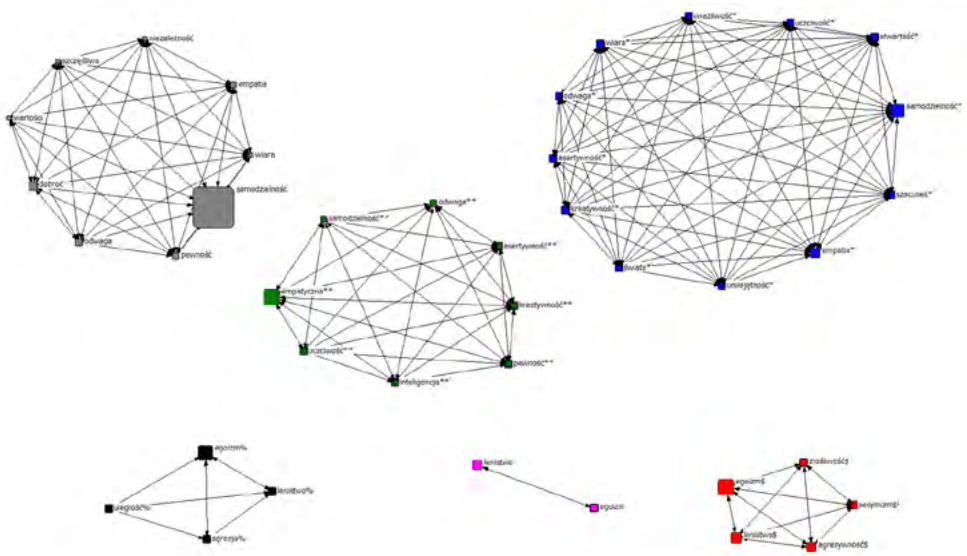


Figure 33.1. Network of nodes, attributes, and relations in the group of parents of girls.

Traits mentioned in the first position – grey, second – blue, third – green, fourth – red, fifth – black, sixth – pink. Legend (most prominent traits): samodzielność – self-reliance, empatyczna – empathetic, umiejętność – skills, pewność – confidence, twórczość – creativity, empatia – empathy, złośliwość – maliciousness, lenistwo – laziness, egoizm – selfishness.

This simple network uses information about two node attributes: the order in which a parental goal is mentioned, and its frequency. Frequency is visualised through node size – it is clear that *independence* constitutes the largest node. Meanwhile, the order in which the trait was mentioned is visualised using colours. In this way, elements mentioned as first, second, third, etc., were linked to one another. The nodes displayed in grey and connected to each other constitute a subnetwork in which *independence*, followed by *kindness*, have the greatest weight (they were the most frequently

mentioned goals), and they were listed first. The goals mentioned in the second position are shown in blue – it is evident that *independence* in the group of parents of girls was also listed in the second position as the most frequently mentioned parental goal.

Thanks to the ability of the network to incorporate multiple attributes, the graphs may become highly complex. In addition to frequency and order, we could also take into account whether a trait was desirable or not. This attribute could be indicated by shape (e.g., desirable traits as stars and undesirable traits as crescents).

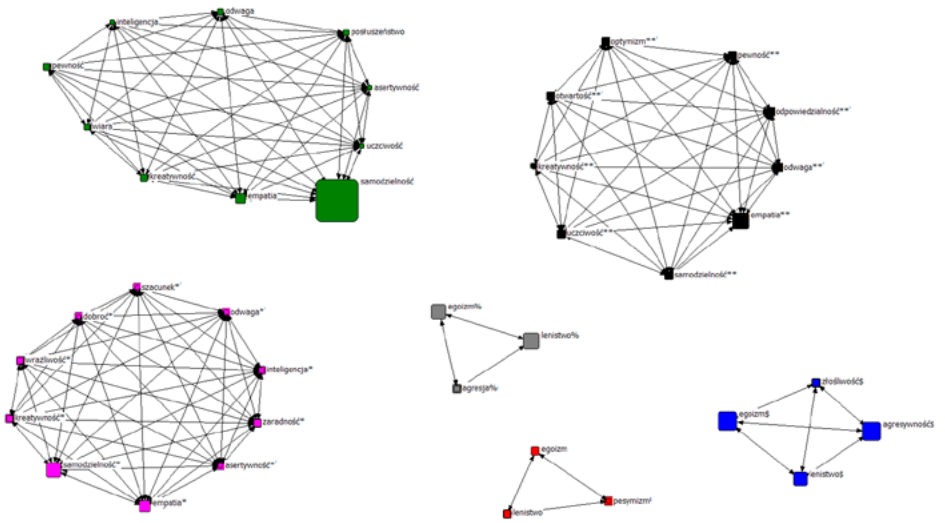


Figure 33.2. Network of nodes, attributes, and relations in the group of parents of boys. Traits mentioned in the first position – green, second – pink, third – black, fourth – blue, fifth – grey, sixth – red.

Legend (most prominent traits): samodzielność – self-reliance, empatia – empathy, empatyczna – empathetic, pewność – confidence, twórczość – creativity, inteligencja – intelligence, złośliwość – maliciousness, lenistwo – laziness, egoizm – selfishness.

In the graph illustrating the nodes in the group of parents of boys (Figure 33.2), we can likewise observe that *independence* is the most desirable parental goal, mentioned both in the first and second position. The way in which the network is presented naturally depends to a large extent on the researcher, which allows one to highlight the attributes that are particularly important in a given analysis.

To summarise, the analysis demonstrated the application of network analytics to examine psychological traits that parents wish to develop in their children. These traits, treated as nodes in the network, were analysed in the context of various attributes such as the child's gender, parental preferences regarding the desirability of the trait, and the frequency with which it was indicated. Directed graphs were used in the study, enabling the representation of hierarchies and parental preferences concerning the traits they wish to foster in their children. This structure was

supported by an attribute matrix containing information on the order of trait mention, desirability, the child's gender, and frequency.

Network analysis methods were employed, such as centrality analysis, which enabled the identification of the most important traits in the network, reflecting their popularity and influence on the modeling of children's personality traits. Clustering allowed for understanding which traits are frequently grouped together by parents, suggesting shared preferences in upbringing. Network density analysis provided insights into the intensity of connections between various traits, which influenced how their importance was perceived.

The analysis reveals that certain traits, such as *independence*, are particularly desirable to parents for both girls and boys, which underscores their universal value in upbringing. The method of presenting network data enables a deeper understanding of the complex interdependencies among various traits, as well as their differentiation based on the child's gender.

The UCINET program was used to process the data and visualise the networks. The use of this tool allows for a comprehensive and multidimensional network analysis, enhancing the quality and depth of research in the field of parental trait analysis. The study demonstrated how Social Network Analysis can be useful in research on parental preferences regarding children's psychological traits, which is relevant from both psychological and educational perspectives. The use of network analysis allows for a deeper understanding of social dynamics and the identification of key factors influencing parental decisions in upbringing.

33.2.2. Second Example: Network Analytics as a Tool for Studying Parenting Dynamics

Network analytics, owing to its ability to model complex structures and the dynamics of relationships, is an indispensable tool in the analysis of social processes occurring over time. The application of this methodology allows not only for the observation of changes in network structure, but also for detailed tracking of the evolution of attributes and relations between nodes. In the context of research on parental goals, a network that incorporates the time factor enables in-depth investigation of how parental preferences change as the child matures and how these changes influence the formation of the child's personality.

In the analysis under discussion, numerical data related to parental goals were transformed into numerical values in accordance with the methods described in the previous chapters. These data were subsequently mapped onto eight metatraits of personality according to the Circumplex Model of Personality, such as Alpha Plus, Beta Plus, Delta Plus, Gamma Plus, and their negative counterparts (Strus et al., 2014).

The personality metatraits identified in the Circumplex Model of Personality represent combinations of the basic dimensions of the Big Five and describe broad, integrative profiles of individual functioning. *Alpha Plus* (Stability) characterises

socially adapted individuals who are conscientious, persistent, and capable of delaying gratification. *Beta Plus* (Plasticity) refers to individuals who are open to change, enthusiastic, and inclined to take initiative and engage in personal development. *Gamma Plus* (Integration) pertains to psychological maturity, well-being, openness, and trust toward others. *Delta Plus* (Restraint) is typical of individuals who are orderly, conventional, and compliant with social norms, yet socially reserved and closed to new experiences.

The negative poles include: *Alpha Minus* (Disinhibition) – antagonistic attitudes, impulsivity, aggression, and disregard for norms; *Beta Minus* (Passivity) – apathy, submission, and fearfulness; *Gamma Minus* (Disharmony) – low mood, lack of energy, and low self-esteem; and *Delta Minus* (Sensation Seeking) – impulsivity, emotional excitability, expansiveness, and hedonism.

Table 33.1 presents a matrix that serves as a representation of nodes in the analysed network. These nodes correspond to the aforementioned personality traits. The matrix includes detailed data on the child's age, the intensity with which a given trait is emphasised by parents, and the classification of the trait as "plus" or "minus". This information allows for a deeper understanding of how individual traits are shaped depending on the stage of the child's development and is essential for analysing the dynamics of changes within the network.

Table 33.1. Distribution of Educational Personality Traits According to the Child's Age

ID	Age group	Intensity	Trait type
AlphaP	three-year-olds	12576	plus
BetaP	three-year-olds	6667	plus
DeltaP	three-year-olds	3788	plus
GammaP	three-year-olds	6970	plus
AlphaM	three-year-olds	14394	minus
BetaM	three-year-olds	2576	minus
DeltaM	three-year-olds	7273	minus
GammaM	three-year-olds	5303	minus
AlphaP*	four-year-olds	11647	plus
BetaP*	four-year-olds	4353	plus
DeltaP*	four-year-olds	4353	plus
GammaP*	four-year-olds	8824	plus
AlphaM*	four-year-olds	14588	minus
BetaM*	four-year-olds	2471	minus
DeltaM*	four-year-olds	8118	minus
GammaM*	four-year-olds	4000	minus
AlphaP**	five-year-olds	15517	plus
BetaP**	five-year-olds	6551	plus
DeltaP**	five-year-olds	5000	plus
GammaP**	five-year-olds	7931	plus
AlphaM**	five-year-olds	13620	minus

ID	Age group	Intensity	Trait type
BetaM**	five-year-olds	1896	minus
DeltaM**	five-year-olds	7241	minus
GammaM**	five-year-olds	5172	minus
AlphaP***	six-year-olds	12898	plus
BetaP***	six-year-olds	2898	plus
DeltaP***	six-year-olds	5797	plus
GammaP***	six-year-olds	7971	plus
AlphaM***	six-year-olds	15942	minus
BetaM***	six-year-olds	4927	minus
DeltaM***	six-year-olds	6521	minus
GammaM***	six-year-olds	1159	minus

To conduct the social network analysis, data concerning the intensity of the listed traits (treated as nodes) and the age categories of the children (also treated as a time axis) were used. Based on age category, the network analysis distinguished subnetworks that represent different developmental phases of children (Figure 33.3). The intensity of traits was visualised through varying node sizes in the network diagrams, allowing for a visual comparison of their dominance across different age categories.

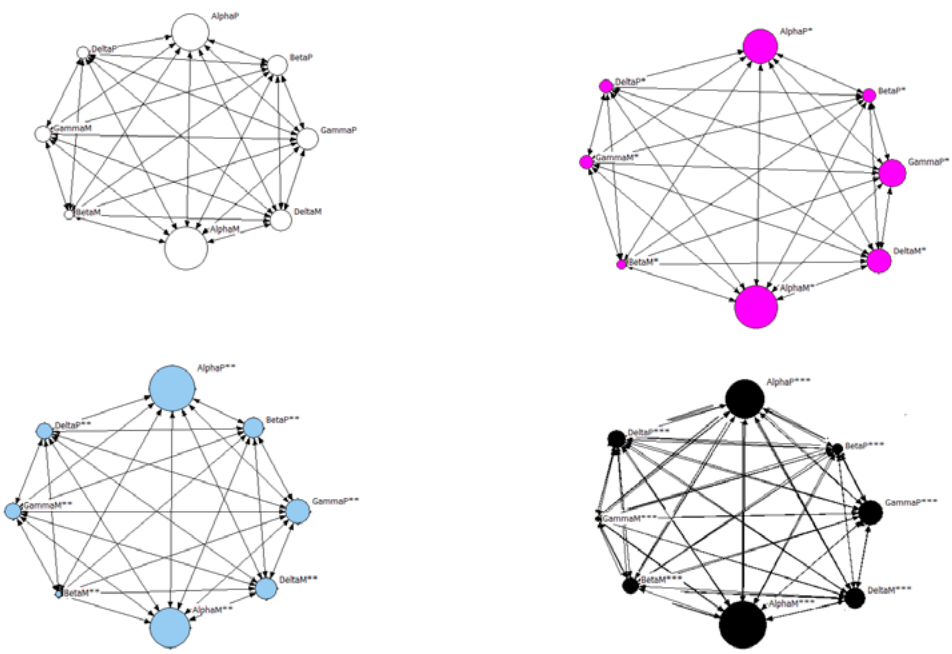


Figure 33.3. Analysis of the Evolution of Personality Traits in Child Development: Age-Based Network Representations (three-year-olds – white, four-year-olds – pink, five-year-olds – red, six-year-olds – black)

Parents of three-year-old children primarily support the development of "plus"

traits such as *Stability* (Alpha Plus), *Plasticity* (Beta Plus), and *Integration* (Gamma Plus). The least promoted trait in this group is *Restraint* (Delta Plus), which appears to be the least developed. At the same time, parents exhibit particular vigilance toward manifestations of *Disinhibition* (Alpha Minus), striving to suppress its development to the greatest extent.

Parents of four-year-old children focus primarily on cultivating the Alpha Plus trait, symbolising concern for the child's social development. This is followed closely by Beta Plus traits, reflecting openness to new experiences, and Gamma Plus traits, representing personality integration. The least emphasised trait remains Delta Plus, associated with restraint.

Compared to three-year-olds, four-year-olds show increased parental interest in developing Gamma Plus at the expense of Beta Plus. A slight increase in emphasis on Delta Plus is also observed. As children grow older, parents of five-year-olds intensify their promotion of Alpha Plus, Delta Plus, Beta Plus, and Gamma Plus traits. Among six-year-olds, the most intensively developed traits are Alpha Plus and Delta Plus, while Beta Plus is relatively less emphasised.

Social network analysis reveals that, as children grow, parents increasingly foster Alpha Plus and Gamma Plus traits, and noticeably strengthen the development of Delta Plus, at the cost of reduced focus on Beta Plus.

As for traits classified as "minus", Alpha Minus receives the most attention as the most undesirable trait that parents seek to suppress. Its status as the least desirable trait remains stable throughout the preschool period. Delta Minus, symbolising unrestrained psychological behaviours, also remains steady but at a lower level across all years. Toward the end of the preschool period, Beta Minus, associated with limitations in openness to experience, becomes increasingly emphasised.

In third position among the most undesirable traits, parents identify Gamma Minus, associated with personality disintegration. Its perceived importance gradually diminishes with the child's age, while by the end of the preschool years, increasing emphasis is placed on counteracting the development of Beta Minus.

Thanks to detailed network analysis, it was possible to examine which personality traits are cultivated by parents at various stages of their child's development, and which traits are suppressed in order to foster socially healthy and well-adapted personalities. Importantly, this analysis—despite its scientific nature—highlights how parents contribute to the construction of society by promoting positive traits and suppressing negative ones. This conscious effort is of crucial importance in shaping future generations.

As part of the discussed section on *Network Analytics as a Tool for Examining Upbringing Dynamics*, advanced Social Network Analysis (SNA) methods were employed to enable in-depth exploration and interpretation of complex patterns of relationships and dynamics within the parental upbringing process. This section focused on the use of graphs, matrices, and advanced network metrics in the context of examining changes in parental preferences as children develop.

The analysis primarily utilised weighted graphs and directed graphs. Weighted

graphs allowed for the representation of differences in the intensity with which personality traits (nodes) were manifested, enabling an assessment of the strength of parental influences across different age groups. Directed graphs, on the other hand, allowed for modeling asymmetries in relationships, which was crucial for understanding the directionality of parental influence on the development of children's traits.

Adjacency matrices were used to document both the presence and the intensity of relationships between nodes in the network, i.e., personality traits. These matrices contained numerical values corresponding to the intensity of parents' emphasis on specific traits, which enabled precise mapping of the network structure and identification of key traits across different child age categories.

A key aspect of the analysis was the assessment of node centrality, which enabled the identification of dominant personality traits in different developmental stages of children. These metrics provided insight into which traits played a central role within the network of parental influences. Additionally, the flow of information within the network, analysed using metrics such as betweenness centrality, yielded data on the dynamics and evolution of traits over time, which is essential for understanding how children's personalities form at different stages of development.

This chapter demonstrated how advanced SNA techniques can be employed to analyse and interpret complex upbringing processes. Through the use of diverse graphs, matrices, and network measures, it was possible not only to observe changes in network structure, but also to gain a deeper understanding of how personality traits are shaped by parental influence. The analysis emphasised the role of SNA as a method that enables in-depth interpretation of interactions and dependencies in social processes, which is of key importance in both methodological and practical contexts.

33.2.3. Third Example: Constructing a Network of Personality Disorders

In this section, devoted to constructing a network of personality disorders, we present the final of three analytical examples. In contrast to the first two, which employed the UCINET software, this example uses advanced artificial intelligence algorithms provided by OpenAI in the form of the GPT language model to illustrate the construction of a network based on shared traits and symptoms of personality disorders.

Personality disorders, as described in numerous studies, often share certain traits (Millon & Davis, 1996), which can be analysed using Social Network Analysis (SNA) methods. Applying this methodology allows for both visual and statistical interpretations that help to elucidate the complex patterns of interdependence among various personality disorders.

Network analysis in clinical psychology enables us to observe how individual

disorders are interconnected through shared symptoms and traits. For example, *sadism*, identified in the literature as sharing common features with negativistic, paranoid, narcissistic, and antisocial personality types, can be connected in the network to each of these disorders through separate edges. Each of these edges symbolises distinct yet frequently overlapping characteristics such as dominance, aggression, or paranoia.

Table 33.2 presents a matrix illustrating the relationships among personality disorders, developed by Szymańska based on descriptions of disorders provided by Davis and Millon (Millon & Davis, 1996). A detailed explanation of these connections and the symptom patterns that are shared across different disorders is available in Szymańska (2024c). Here, only the matrix is shown, indicating which disorders are linked through at least one similar symptom.

The adjacency matrix depicting the relationships between personality disorders was generated by advanced artificial intelligence algorithms provided by OpenAI via the GPT language model. This process was based on an analysis of the literature presented by Szymańska, allowing for a detailed understanding of the connections between various types of disorders (Szymańska, 2024c). The matrix, organised as a square table, includes individual personality disorders in both its rows and columns. At the intersection of a row and column, the symbol "X" indicates the presence of shared traits or symptoms between the analysed disorders, while empty cells suggest no direct connection. For instance, the *sadistic* disorder (first column) shares common features with *negativistic*, *paranoid*, *narcissistic*, and *antisocial* disorders.

The matrix structure is symmetrical with respect to the main diagonal, which is typical for undirected graphs and implies that each pair of disorders is mutually and equally connected. The absence of markings along the main diagonal, where each disorder intersects with itself, highlights the irrelevance of self-relations in the context of analysing shared symptoms.

This adjacency matrix constitutes a valuable research tool in the fields of psychology and psychiatry, enabling the identification of key patterns within networks of personality disorders. It facilitates the recognition of groups of disorders that are strongly interconnected through shared traits, contributing to a deeper understanding of the mechanisms underlying their comorbidity and impact on individual functioning. This approach has significant implications for diagnosis, treatment, and the development of therapeutic interventions that effectively address the complexity of issues faced by individuals with personality disorders.

Based on this matrix, artificial intelligence algorithms also generated a corresponding SNA network, the result of which is illustrated in Figure 33.4**.** This visualisation allows for an understanding of how various personality disorders are interconnected through shared symptoms, offering insight into the complexity of their mutual relationships.

Table 33.2. Matrix of Associations Between Personality Disorders

	Sadistic	Negativistic	Paranoid	Narcissistic	Antisocial	Histrionic	Borderline	Dependent	Masochistic	Obsessive--compulsive	Avoidant	Schizotypal	Depressive	Schizoid
Sadistic	–	X	X	X	X									
Negativistic	X	–	X	X	X			X	X			X		
Paranoid	X	X	–	X	X		X	X		X	X			
Narcissistic	X	X	X	–	X	X	X							
Antisocial	X	X	X	X	–	X	X							
Histrionic				X	X	–	X	X						
Borderline			X	X	X	X	–	X	X	X		X	X	X
Dependent		X	X			X	X	–	X	X	X			
Masochistic			X	X			X	X	–	X	X		X	
Obsessive--compulsive			X				X	X	X	–	X	X	X	X

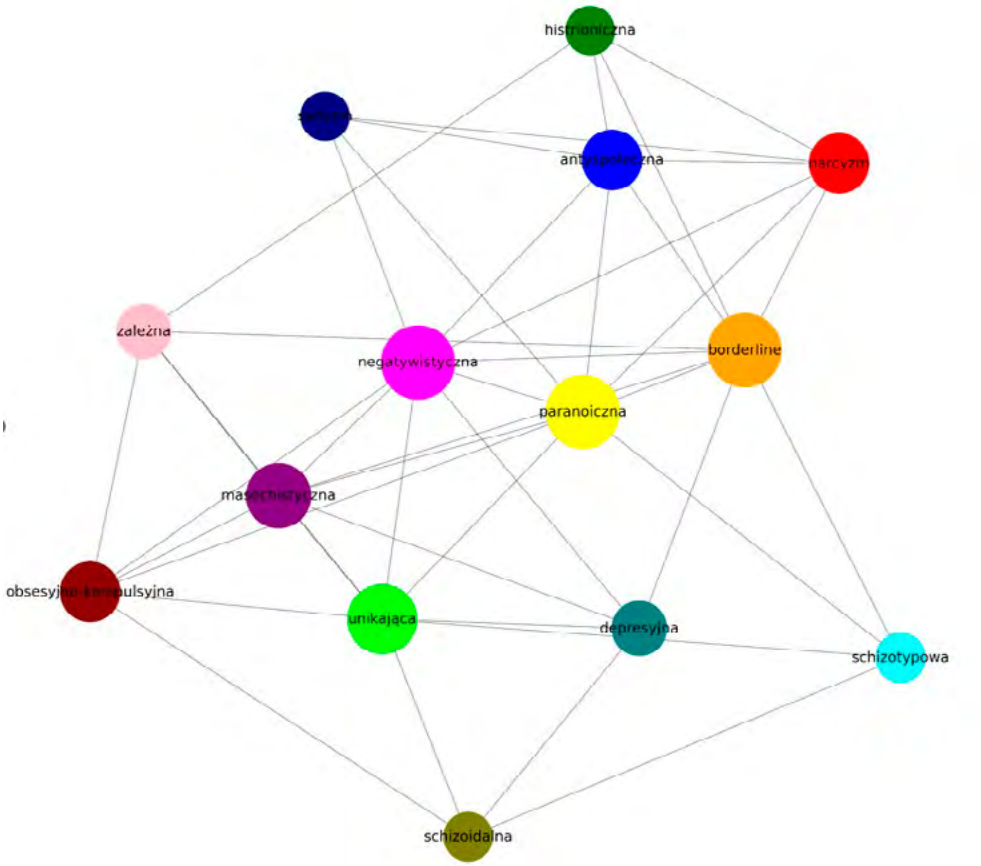


Figure 33.4. Graph of connections among personality disorders based on shared symptoms

Legend: Sadistic – sadystyczna, Negativistic – negatywistyczna, Paranoid – paranoiczna, Narcissistic – narcystyczna, Antisocial – antyspołeczna, Histrionic – histrioniczna, Borderline – borderline, Dependent – zależna, Masochistic – masochistyczna, Obsessive-compulsive – obsesyjno-kompulsyjna, Avoidant – unikająca, Schizotypal – schizotypowa, Depressive – depresyjna, Schizoid – schizoidalna

The graph presented illustrates a complex network of connections among various personality disorders, where each node represents a distinct disorder. These nodes vary in size, reflecting the number of shared symptoms that each disorder has with others. Larger nodes—such as *negativistic*, *paranoid*, and *borderline* disorders—indicate that these disorders share a broader range of traits with others, making them key points in this network structure. These disorders are characterised by the following shared features.

Negativistic disorders are marked by a tendency to perceive the environment negatively, which may overlap with fear of rejection or negative evaluation observed in *avoidant* and *dependent* disorders. *Paranoid* disorders, manifested in general

distrust and suspicion, may overlap with interpersonal fears typical of *avoidant* disorders. *Borderline* disorders, with their intense yet unstable emotional relationships, may share traits such as impulsivity and fear of abandonment, characteristic of *histrionic* and *dependent* disorders.

Smaller nodes—such as *schizoid* or *obsessive-compulsive* disorders—while connected by fewer shared symptoms, are nonetheless significant, as they indicate more specific links that may be crucial in particular clinical contexts.

The graph allows us to observe which disorders are most closely related through direct proximity within the network. For example, the closeness of the *borderline* and *histrionic* nodes suggests frequently co-occurring features such as high emotionality and a strong desire to be the centre of attention. Network analysis presented in this way is particularly helpful in understanding how different disorders may jointly influence patient functioning, which is fundamental to effective diagnosis and treatment.

The analysis of the adjacency matrix and the resulting graph of connections among personality disorders, developed using advanced artificial intelligence algorithms provided by OpenAI in the GPT language model, confirms their usefulness in scientific social network analysis (SNA). Moreover, the consistency between results obtained via artificial intelligence algorithms and those published from analyses conducted using the R software—based on the same dataset—further highlights the effectiveness of these tools in SNA (Szymańska, 2024c). This dual confirmation not only strengthens the credibility of the methods but also indicates that modern AI technologies can be employed just as effectively in constructing SNA networks as traditional approaches.

In the research on personality disorders, the artificial intelligence algorithms provided by OpenAI in the GPT language model demonstrated an exceptional ability to analyse and synthesise complex textual data. They were applied to process a comprehensive article describing the relationships among various personality disorders, which contained detailed verbal descriptions spanning several pages. These algorithms, operating autonomously, not only absorbed the textual content but were also able to construct an adjacency matrix and a corresponding SNA graph based on it. This achievement illustrates not only the technological sophistication of the tools employed, but also their practical value in the context of clinical psychology, enabling a deeper and more systematic understanding of the structure and dynamics of relationships among personality disorders. As a result, these algorithms not only support traditional research methods but also open new possibilities for scientific network analysis that would be difficult to achieve using conventional approaches.

Summary

In this chapter dedicated to practical workshops and tutorials, three distinct examples of applying Social Network Analysis (SNA) in psychological research were presented. Each example illustrates a different methodology and the use of various

analytical tools, allowing for a thorough understanding of the potential of this technique in both clinical and upbringing psychology.

The first example focused on analysing psychological traits that parents wish to cultivate in their children, demonstrating the use of the UCINET software to visualise preferences depending on the child's gender and the way these preferences are perceived by parents. The second example showed how SNA can be applied to understand upbringing dynamics by analysing how personality traits are shaped by parents at different stages of child development, also using UCINET. The third example highlighted the advanced capabilities of artificial intelligence algorithms in the GPT language model from OpenAI, whose application enabled the autonomous synthesis and analysis of complex textual data to generate an adjacency matrix and a corresponding SNA graph illustrating the complex interrelations and shared features among different personality disorders.

Although key techniques and tools were discussed, this chapter does not exhaust all possible SNA methods. For instance, advanced techniques such as community detection or dynamic network analysis were not addressed, though they may offer deeper insights into complex social processes. These techniques may serve as a direction for further research, expanding the scope and depth of SNA applications in psychology.

The examples presented not only enrich the psychological literature but also open new research avenues, emphasising the importance of further development and adaptation of SNA methods. This chapter underscores how social network analysis can be used to understand complex patterns and dynamics in psychology, and how diverse tools and methods may be applied depending on the specificity of the research problem—an essential factor in advancing psychological modeling through modern technologies.

PART VII

Application of Artificial Intelligence Algorithms in Psychometrics

CHAPTER 34

Introduction – Definition, Contemporary Approaches to Psychometrics, and the Role of Artificial Intelligence in Psychometrics

Measurement in psychology plays a key role in both scientific research and diagnostic practice. According to the American Psychological Association (APA), measurement is defined as the systematic assignment of numerical values to specified psychological characteristics based on established rules (American Psychological Association, 2020). This process enables the quantification of psychological constructs, which is foundational in both research and clinical practice. Psychological measurement is thus inherently linked to the idea of precisely assessing such phenomena as intelligence, personality, emotions, or cognitive abilities, allowing for their comparison, analysis, and interpretation of results.

The fundamental measurement tool in psychology is the psychological test. There are many definitions of a psychological test, each highlighting its various aspects. According to APA standards, a psychological test is a diagnostic procedure that may take the form of a set of tasks or questions. The purpose of the test is to elicit specific behaviors that make it possible to obtain results reflecting desired psychological attributes such as reliability and validity (Hornowska, 2003). Psychological tests are thus instruments that allow for accurate and repeatable measurements of individuals' psychological traits.

Similarly, Chojnowski defines a psychological test as an instrument consisting of a set of questions or situational simulations designed to examine the psychological properties of an individual or group. He emphasizes that these tests elicit specific verbal or non-verbal responses that are representative of the behaviors of the

examined person or group (Aranowska, 2005; Hornowska, 2003). In turn, Anastasi and Urbina describe a test as a standardized tool that allows for the collection of samples of human behavior (Anastasi & Urbina, 1999).

Across all definitions of the psychological test—those cited by the APA, as well as by Chojnowski and by Anastasi and Urbina—there emerges a common element that can be considered essential to understanding the nature of testing. This is the function of the test as a means of creating standardized stimulus situations intended to elicit specific responses from the individuals being assessed. These responses are then analyzed with regard to the psychological traits of interest to the researcher.

The psychological test is therefore an instrument that functions in a systematic and controlled manner. All participants are exposed to the same stimuli under identical testing conditions, which allows one to assume that any differences in their responses result primarily from their individual personality traits, rather than from external confounding factors. The standardization of testing procedures is essential, as it ensures that the variability in test results can be attributed to actual differences between individuals rather than to fluctuations in testing conditions (Anastasi & Urbina, 1999; Aranowska, 2005; Hornowska, 2003).

This common element found in all definitions highlights a fundamental principle of psychometrics**: psychological measurement is based on the creation of conditions under which it is possible to assess individual differences in a precise and reliable manner.** As a result, psychological tests are capable of providing valid data that can be used in diagnostics, personnel selection, and scientific research.

Further developing this idea, it becomes evident that the creation of standardized stimulus situations is a necessary but not sufficient condition for obtaining reliable measurement results. A key aspect is also the control over response variability so that it reflects only individual differences in personality traits, rather than, for example, differences in the interpretation of questions or other confounding variables. Therefore, the standardization of a test encompasses not only the testing procedure itself but also the construction of test items, which must be carefully selected in order to accurately measure the traits of interest to the researcher.

In order for a psychological test to be recognized as a high-quality diagnostic tool, it must meet specific criteria that define its psychometric soundness. These criteria concern the psychometric properties of the test, such as validity, reliability, standardization, norming, and objectivity (Anastasi & Urbina, 1999; Hornowska, 2003).

One of the key aspects of test soundness is validity—that is, the degree to which the test measures what it is intended to measure (Anastasi & Urbina, 1999). Validity is not only a matter of the instrument itself but also of the way in which its results are interpreted. The contemporary approach to validity, in accordance with the standards of the American Psychological Association (APA), emphasizes that validity is no longer attributed solely to the test itself but rather to the interpretation of the results it produces (Rytel, 2021). This means that the assessment of validity must include an analysis of how the test results are used in a diagnostic context and what conclusions are drawn from them.

In the past, validity was attributed directly to the measurement tool—one would ask whether a given test measures a specific psychological trait (Anastasi & Urbina, 1999; Hornowska, 2003). Today, the approach to validity has evolved toward a more dynamic interpretation. Validity is now understood as a feature that depends not only on the test itself but also on the context in which the results are interpreted and on the goals that are intended to be achieved through their use. This shift in thinking underscores the importance of the diagnostic context, which leads us to the next key element—psychological diagnosis.

Psychological diagnosis is a process in which the results of psychological tests are integrated with other sources of information in order to gain a comprehensive understanding of the examined individual. According to the definition, psychological diagnosis involves the collection, analysis, and interpretation of data concerning an individual's psychological functioning in clinical, educational, or occupational contexts (Anastasi & Urbina, 1999). Psychological tests, constituting an essential part of the diagnostic process, provide valuable information that aids in formulating accurate and useful diagnostic conclusions.

Contemporary approaches to psychological diagnosis therefore emphasize not only the test results but also the way they are interpreted, which highlights the growing importance of context and the complexity of the diagnostic process.

At this point, we reach the moment where traditional frameworks of psychometrics intersect with modern technologies, and artificial intelligence (AI) begins to play a key role in the further development of the field. AI introduces new opportunities that not only facilitate the precise design and analysis of psychological tests but also open new perspectives for the adaptability and personalization of diagnostic tools. The role of artificial intelligence in contemporary psychometrics is becoming increasingly significant, although it remains a developing area that entails both potential and challenges. AI finds application in various aspects of psychometrics, and its impact can be observed across several key domains.

First and foremost, AI supports the process of creating and calibrating psychological tests. Thanks to advanced machine learning algorithms, it is possible to analyze large datasets and identify patterns that may be difficult to detect using traditional statistical methods. These algorithms not only assist in analyzing existing test items but can also substantially support the construction of new tests by suggesting appropriate psychometric indicators.

As Szymańska (2024b) notes, artificial intelligence may play an important role in the process of constructing psychological tests by assisting in the selection of test items and optimizing them in terms of validity and reliability. In this context, particularly noteworthy is the application of language models such as BERT and GPT, which—due to their contextual analysis of text—can raise the quality standards of diagnostic tools. Their capacity to model meanings and semantic dependencies allows for a better alignment of question content with the measured construct, thereby increasing the validity of psychometric instruments. AI can also identify ambiguities in formulations and suggest improvements that enhance the

precision of tests. Naturally, the entire process requires human oversight, as the final decision regarding item selection and interpretation should remain in the hands of experts.

Artificial intelligence also plays an important role in the analysis of psychometric data. Traditional analytical methods such as structural equation modeling (SEM) or factor analysis are being augmented by AI techniques, such as artificial neural network analysis (Szymańska, 2018, 2019). These techniques enable more precise modeling of complex relationships between variables and allow for better handling of nonlinearity in psychometric data.

Moreover, AI can support the process of adapting psychological tests for different cultural and linguistic groups. By analyzing large datasets, AI algorithms can identify cultural differences that influence test outcomes and indicate elements that may lead to cultural bias. In this context, language models such as BERT and GPT are particularly relevant. Their capacity for semantic analysis and contextual understanding allows them to assist in the translation and cultural adaptation of test items, thanks to their extensive knowledge of cultural specificity and the structure of a given language.

Traditional methods of test adaptation require precise translation and an analysis of semantic equivalence between languages, which often involves difficulties in preserving the original meaning of test items. Language models can significantly improve this process by analyzing differences in word connotations and suggesting the most appropriate formulations that better capture the intended construct in a given culture. Additionally, AI can detect ambiguities in translations and indicate potential areas where a given version of the test might distort results due to specific cultural norms or differences in the understanding of psychological concepts.

As a result, artificial intelligence not only improves the technical process of test adaptation but also enhances test quality by increasing both validity and diagnostic fairness across various respondent groups.

One of the more innovative applications of AI in psychometrics is the development of intelligent systems that support psychological diagnosis, such as expert systems (Szymańska, 2019, 2024b). These systems, based on the analysis of psychological test data, can support psychologists in making diagnostic decisions by offering suggested diagnoses or interpretations of results. Such solutions can increase the accuracy of diagnosis, especially in complex cases or those difficult to assess using traditional methods.

One of the key challenges in psychometrics is optimizing the diagnostic process in such a way that tests remain reliable and valid, but without overburdening the respondent with excessive length. Traditional methods rely on static sets of questions that do not take into account the individual's real-time responses. As a consequence, tests may contain many items that are unnecessary for a given person, resulting in prolonged diagnostic time and a potential decrease in respondent motivation. Expert systems offer a response to this issue by enabling more flexible and precise testing.

The chapter titled *"Expert Systems as Intelligent Tests"* presents the concept of intelligent tests based on artificial intelligence algorithms. It describes how expert systems can dynamically guide the testing process by intelligently selecting subsequent questions based on the respondent's earlier answers. Such a system is capable of eliminating redundant questions and focusing on those that are crucial for the diagnosis.

Additionally, this chapter will address the role of artificial intelligence in monitoring the reliability of measurement through the implementation of algorithms capable of real-time assessment of the consistency and validity of obtained results. The result is an intelligent, adaptive psychometric diagnostic system that can significantly enhance the efficiency and precision of measurement.

It should be emphasized, however, that despite its enormous potential, the application of AI in psychometrics also brings challenges. One of these is the necessity to ensure that AI algorithms operate in a transparent and comprehensible manner for users. Furthermore, it is essential to guarantee that the results generated by AI-based systems comply with psychological ethical standards and do not lead to discrimination or distortion of results.

Particular attention should be paid to contemporary methods of psychological trait profiling, especially in the context of multidimensional questionnaires and the dynamic alignment of individual profiles with diagnostic patterns. The concept of psychological profiling has functioned in psychology for decades and is based on the interpretation of psychological self-portraits created on the basis of test results. Multidimensional questionnaires allow for the presentation of results in the form of trait profiles, which are used in clinical diagnosis, occupational psychology, and the assessment of social competencies.

However, the development of artificial intelligence opens entirely new possibilities in this area. Thanks to machine learning algorithms, it is possible to automatically match test results to reference profiles characteristic of specific groups, such as individuals with personality disorders, distinctive thinking and behavioral patterns, or predispositions to particular behaviors.

The potential of this technology goes far beyond traditional psychometrics. AI can assist psychologists in developing new profiling methodologies by analyzing psychometric patterns and identifying subtle relationships that were previously difficult to detect. Initial attempts at such applications are already being undertaken, and the continued development of this technology will allow for more precise identification of characteristic features of specific psychological profiles—for example, psychopathy in the context of the Big Five model or other personality frameworks (Szymańska, 2025b).

It is worth emphasizing that the application of AI in profiling is not limited to automating the diagnostic process. It may also contribute to the implementation of new models in clinical, forensic, or organizational practice. With the help of artificial intelligence, it will become possible to isolate specific configurations of traits that were previously difficult to define clearly, as well as to predict how particular constellations of traits may influence individual behavior across various contexts.

The application of artificial intelligence in psychological profiling represents one of the most promising yet ethically demanding challenges in contemporary psychometrics. The development of this technology and its implications will be discussed in detail in the chapter *"Artificial Intelligence in Profiling Psychological Traits: New Approaches in Diagnosis and Psychometric Modeling",* where both the opportunities and limitations, as well as the potential risks of implementing AI systems in this area, will be presented.

Artificial intelligence opens new possibilities for verifying psychometric models on an unprecedented scale. It is well known that psychological measurement is complex, and some models are particularly difficult to verify empirically—especially when the measured trait has a complex, multidimensional nature, such as intelligence, temperament, or personality.

Psychology has developed many theoretical models, yet some of them pose exceptional difficulties in the verification process. In this book, we have already discussed circular models and the available methods for testing them. But what about even more complex structures—those that encompass more than eight or ten traits? The verification of such complex models, in light of contemporary psychometrics, is nearly impossible due to limitations related to dimensionality reduction and the mathematical representation of data. This issue will be addressed in the chapter titled *"The Limits of Classical Psychometrics: Artificial Intelligence and Multidimensional Models".*

Yet this does not exhaust the topic. What if psychological models not only describe multidimensional traits but even assume the existence of additional mathematical dimensions—ones that classical psychometric models do not account for? Examples include personality or intelligence models with three-dimensional structures that go beyond standard two-dimensional mathematical spaces, such as the circular model.

Artificial intelligence can offer new opportunities for verifying both exceptionally complex models and those that assume geometric structures beyond classical approaches. This topic will be examined in the chapter titled *"Psychometric Models in Higher Dimensions: How Artificial Intelligence Can Expand the Measurement Space".*

In summary, artificial intelligence has the potential to revolutionize psychometrics by introducing new tools and methods for data analysis. However, its full implementation requires further research, development, and close monitoring of processes to ensure its use remains responsible and ethical.

This part of the book will examine how artificial intelligence can support psychometrics, particularly in areas such as improving reliability, enhancing validity, and constructing psychological scales. The future of AI in psychometrics will also be considered.

To conclude, we turn to a key question concerning the future of measurement in psychology: in the era of artificial intelligence, is classical psychometrics capable of meeting the challenges of contemporary science? Are generalizability theory and other existing measurement models still sufficient in light of an evolving research and technological landscape?

This is not merely a matter of technical improvement in measurement but rather a fundamental question about its very nature. Is the classical reference to population norms—evaluating the individual by their similarity to the rest of society—still adequate? Or do we need a new measurement paradigm, one that embraces a more dynamic, contextual, and systemic approach to the assessment of psychological processes?

Moreover, does psychology itself—the discipline that psychometrics serves—not currently face the problem of excessive conceptual fragmentation? Classical theoretical frameworks often appear scattered and inconsistent, making their integration an increasingly difficult task. Perhaps we need a different approach—one that not only enables a new description of known processes but also allows for their verification and deeper understanding.

In this context, psychocybernetics emerges as a new direction—one that, like artificial intelligence, aims at the systemic modeling of psychological processes. However, unlike AI, which is based on predictive and statistical algorithms, psychocybernetics focuses on the structure of control and regulation of behavior.

It is not artificial intelligence that poses the key questions about psychological mechanisms—it is psychocybernetics that may allow us to understand them more fully by redefining existing psychological paradigms. In the face of AI's growing influence in science, it may be the systems approach that proves to be the missing piece in modeling the human mind.

These considerations will be addressed in the final chapter of this part of the book, titled *"New Frontiers in Psychometrics: Artificial Intelligence, Psychocybernetics, and the Redefinition of Psychometric Paradigms"*. In that chapter, we will explore alternative conceptions of normativity and the possibilities for integrating psychological models—including models of personality and the description of psychological processes—through the lens of psychocybernetics.

CHAPTER 35

Reliability and Validity of Psychometric Tests Supported by Artificial Intelligence

35.1. Reliability of Psychometric Tests Supported by Artificial Intelligence

Reliability is one of the key concepts in psychometrics, defining the degree of consistency and repeatability of results obtained through a psychological test. Reliability means that a diagnostic tool yields stable, trustworthy results that remain consistent regardless of circumstances such as the time of testing or the person administering the test (Aranowska, 2005). In practice, this means that if the same test is administered repeatedly under similar conditions, the results obtained by the examined individual should be comparable. High reliability is therefore essential for a psychological test to be recognized as a credible and useful tool for diagnosing psychological traits (Aranowska & Rytel, 2013).

In the context of reliability, researchers apply various methods to ensure that a psychological test meets these standards. One method of assessing reliability is *test-retest reliability*, which involves administering the same test on two different occasions. If the results from both sessions are similar, the instrument can be considered to exhibit *absolute stability*. In this view, reliability reflects the test's ability to deliver stable results regardless of the passage of time, which is particularly important when the aim is to assess how a given psychological state or trait changes—or remains unchanged—over time. Conversely, when a test is administered repeatedly over a short time span, we refer to the instrument's *consistency*. This method evaluates whether the test is coherent and repeatable over a brief interval (Brzeziński, 1996; Choynowski, 1971, as cited in Hornowska, 2003).

A second important method involves comparing results obtained from *parallel tests*, that is, alternative versions of the same instrument designed to measure the same psychological construct. This is referred to as *inter-test equivalence*. Inter-test equivalence assesses whether different forms of the same test yield comparable results, which is especially important in situations where administering the exact same test repeatedly could lead to learning effects or other distortions. When a test is administered multiple times using different versions, we assess its *inter-test equivalence*. When the same version of the test is administered at different times, we assess the *relative stability* of the instrument (Hornowska, 2003).

Another approach to assessing reliability is the *split-half reliability method*, which involves dividing the test into two parts and comparing the results obtained from each half. This division allows researchers to assess the extent to which different segments of the test consistently measure the same psychological trait. It is particularly useful when examining whether all parts of the test are evenly constructed and equally effective in measuring the intended construct (Choynowski, 1971, as cited in Hornowska, 2003).

Internal consistency, often assessed using Cronbach's alpha coefficient, refers to the degree to which individual test items are correlated with one another and how well they function together to measure the same psychological trait. Although Cronbach's alpha is one of the most widely used coefficients, more recent methods, such as McDonald's omega, provide more precise estimates of internal consistency by better accounting for the structure of the instrument. These methods aim to overcome the limitations of Cronbach's alpha, offering a more accurate and comprehensive approach to reliability assessment (Hornowska, 2003).

In contemporary research, there is growing emphasis on the fact that Cronbach's alpha is not always an adequate measure of reliability, especially in the context of psychological studies (Aranowska, 2005). Psychometric literature has highlighted that this index is particularly sensitive to the number of items in a test and to their content similarity. Too often, the value of alpha increases not because of genuine scale consistency but due to artificially increasing the number of items or excessive item redundancy.

Based on my own observations and simulations, I have found that Cronbach's alpha may lead to the rejection of those items that best differentiate respondents. In experiments in which I compared the value of alpha with other indicators, such as the point-biserial correlation between an item and the total score (for binary data), I observed that items with the highest discriminating power—as indicated by high point-biserial correlation—were also the ones most burdensome for alpha and often eliminated by it. Such phenomena may partly explain why psychometric literature increasingly raises critical voices against Cronbach's alpha, pointing out its limitations in accurately assessing the diagnostic quality of test items.

Each of these methods is aimed at ensuring that a diagnostic instrument is not only theoretically sound but also practically useful—delivering results that are both consistent and replicable. Test reliability is thus the foundation upon which all psychological diagnostics rests.

Modern methods for supporting the assessment of psychological test reliability can benefit significantly from the use of artificial intelligence (AI). Most notably, AI can play a crucial role with regard to two reliability assessment methods: *inter-test equivalence* and *split-half reliability*, that is, *parallel-form reliability* and *within-test equivalence*.

The split-half method (also referred to as *inter-part* or *inter-item equivalence*), known as *split-half reliability*, involves dividing a test into two or more parts and comparing the results obtained from each of them. Artificial intelligence can significantly enhance this process by dividing the dataset into several parts and using advanced algorithms to predict outcomes based on one part of the test, then verifying these predictions against the remaining parts.

For example, consider a psychological test composed of three sections. AI could analyse a respondent's performance on the first section of the test and then, using artificial neural networks, predict the results the individual should achieve on the second and third sections. If the person achieved 70% of the maximum score on the first section, the AI would predict that the scores on the remaining sections should also hover around that value—provided the test is reliable. The AI then compares these predictions with the actual scores, assessing the consistency across different parts of the test. This type of analysis enables dynamic assessment of split-half and inter-item equivalence, which, in traditional methods, can be significantly more difficult and less precise.

Similarly, artificial intelligence can support the testing of parallel-form reliability, which involves comparing results obtained from different but equivalent versions of the same test. In this case, AI can predict the results on one version of the test based on the results from another version.

For instance, if a respondent scored 80% of the maximum points on one version of the test, the AI may predict that the score on the parallel version should also be approximately 80%, assuming the tests are equivalent. The AI can then compare the predicted results with the actual outcomes to determine whether convergence exists between the tests. If so, this confirms that the tests are equivalent and that the diagnostic tool is reliable.

An example of the use of artificial intelligence in psychometrics is the study by Wang and colleagues (2023), who refer to the emerging area known as *Psychometric AI*. The authors applied machine learning algorithms to the measurement of psychological traits, demonstrating the potential of AI as a tool supporting the assessment of the emotional intelligence construct (Wang et al., 2023). Their approach aligns with the broader trend of seeking new, objective, and automated methods of psychological measurement.

The application of AI in the reliability assessment of psychological tests appears to be an innovative and promising approach that warrants further research and development. In particular, artificial intelligence algorithms such as Random Forest, SVM, and other machine learning techniques described in this book may potentially support the analysis of psychological test reliability, offering new possibilities for modeling and precise analytical tools that could complement classical psychometric approaches.

The use of AI in reliability testing brings numerous benefits. Above all, it enables precise and rapid processing of large datasets, which is particularly important when tests consist of many items or when multiple versions of a test are involved. AI is also capable of identifying subtle patterns in data that may remain undetected using traditional methods, thereby further increasing the accuracy and credibility of reliability testing.

Thanks to advanced machine learning algorithms, AI can predict how different parts of a test or various versions of the same test will align with one another, allowing for more dynamic and accurate reliability assessment. In this way, AI becomes not only a supportive tool but one that is actively revolutionising traditional methods for evaluating the reliability of psychological tests.

However, it should be noted that the use of AI for such purposes is still in the developmental phase, and there are not yet widely accepted standards or guidelines regulating these practices in psychometrics. Articles addressing the future of psychometrics with AI emphasise both the immense potential and the challenges associated with integrating these technologies—particularly in the context of ensuring the accuracy and reliability of psychological outcomes (Szymańska, 2024b).

35.2. Validity of Psychometric Tests Supported by Artificial Intelligence

Validity is one of the key concepts in psychometrics and refers to the extent to which a test measures what it is intended to measure. Validity indicates the degree to which the results obtained through a given test are appropriate and useful for the intended measurement purpose (Anastasi & Urbina, 1999; Aranowska, 2005; Hornowska, 2003; Rytel, 2021). It is a fundamental characteristic of any diagnostic tool that directly influences its credibility and practical utility.

In contrast to reliability, which pertains to the *consistency* and repeatability of test results, validity focuses on the quality of measurement in the context of what the test is actually supposed to measure. A valid test is therefore always reliable, as its results are stable and consistent; however, a reliable test is not necessarily valid if it measures something other than what it is intended to assess.

Calibration, in the context of psychometric testing, refers to the process of adjusting a test so that its results are as adequate as possible for the intended measurement purpose. In psychometric practice, calibration may involve adapting test items, scales, or scoring keys in such a way that the results most accurately reflect the constructs being measured. Calibration is thus one of the tools used to improve a test's validity.

Calibration and validity are closely interrelated. Proper calibration of a test can significantly enhance its validity, thereby increasing confidence that the test results genuinely reflect the construct under investigation. At the same time, validity presupposes measurement reliability—a test that is properly calibrated and valid will also be reliable, as it will yield stable and trustworthy results.

In the past, validity was understood as a characteristic attributed directly to the test—it was assessed whether the test itself was valid. Today, in line with the definition proposed by the American Psychological Association (APA), validity is attributed more to the interpretation of test results than to the test itself. This means that validity pertains to the extent to which the results obtained are adequate for a specific context or decision, rather than being solely a property of the diagnostic instrument. A thorough analysis of the changes in the understanding of this concept is provided by Jolanta Rytel (Rytel, 2021).

The contemporary approach emphasises that validity is a dynamic process that depends on the context in which a test is used. Modern psychometric theory has moved away from the former classification into separate types of validity—such as content, criterion, and construct validity. Instead, current thinking refers to various *aspects* of validity, which better reflect the complex and integrated nature of this concept (Messick, 1989, as cited in Rytel, 2021). Rather than treating these aspects as distinct types of validity, the modern approach highlights that they represent different dimensions of the same property—namely, measurement validity.

The content aspect of validity refers to the extent to which the test content covers all relevant elements of the construct being measured. This is a fundamental dimension that ensures the test genuinely reflects what it is intended to measure. While content validity was once treated as an independent type of validity, it is now regarded as one of several aspects that operate together to ensure the overall validity of the test. This aspect is particularly important during the test construction phase, where it is essential that each item accurately represents the phenomenon under investigation.

The criterion-related aspect of validity concerns the degree to which test scores correlate with external criteria that are recognised as measures of the same or related constructs. Previously, criterion validity was understood as the test's ability to predict future outcomes (predictive validity) or to align with other tests administered concurrently (concurrent validity). Today, however, this is seen as part of a broader understanding of validity, in which the evaluation of associations with external criteria constitutes one aspect supporting the overall assessment of test validity.

The theoretical aspect of validity (formerly referred to as construct validity) pertains to the extent to which the test measures theoretical constructs in accordance with theoretical assumptions. This aspect is crucial, as psychometric research is grounded in theories that assume test results reflect specific psychological traits or phenomena. Theoretical validity requires a series of empirical studies to confirm that the test indeed measures what it is theoretically intended to measure.

In summary, the contemporary approach to validity in psychometrics assumes that validity is a multidimensional construct encompassing various aspects that work together to support the comprehensive evaluation of a test's validity. As Rytel (2021) emphasises, such an approach allows for a more precise and comprehensive assessment of psychometric tools, taking into account the multiple dimensions of measurement that collectively influence test quality.

35.3. Factors Affecting the Validity and Reliability of Measurement

The length of a psychometric test is one of the key factors that can significantly influence both the validity and reliability of measurement. Although longer scales may provide more data and theoretically offer a better reflection of the complexity of the construct being measured, they also have notable drawbacks.

The longer the scale, the greater the risk that respondents will experience fatigue, boredom, or frustration while answering. This, in turn, leads to reduced concentration, which may result in responses that are less considered, more random, or disorganised. Consequently, the results of such a test may be less valid, as they do not reflect the actual traits or states of the individuals, but rather their increasing fatigue and declining motivation.

Long scales can also affect measurement reliability, which is directly linked to increased variability in the latter parts of the test, when respondents may stop engaging with the items in the same way as they did at the beginning. In such cases, the results may become inconsistent, thereby lowering the overall quality of the test.

Control scales designed to detect random or inconsistent responses may help identify problems arising from excessive scale length, but they are not always sufficiently effective. In practice, when a test scale is too long, control scales may indicate high levels of randomness or inconsistency in responses, rendering the entire test unreliable. In such situations, the validity of the test is undermined, and the results become diagnostically unusable.

In this context, artificial intelligence (AI) may play a crucial role in optimising test scale length, ensuring that the testing time is short enough to minimise respondent fatigue, yet long enough to ensure both the validity and reliability of measurement.

Using advanced algorithms, AI can simulate different test lengths, examining how the validity and reliability of the test vary depending on the number of items. This may lead to the development of shorter, more efficient tests that are less burdensome for respondents while still providing reliable results.

In summary, the length of a psychometric test scale is a critical factor affecting its validity and reliability. The application of artificial intelligence may assist in optimising the duration of assessment, thereby improving measurement quality by minimising the negative effects associated with prolonged respondent engagement. These issues will be further discussed in the chapter entitled *"Expert Systems as Intelligent Tests"*.

An example of the effective application of a mathematical algorithm to reduce the number of items in a psychological scale is the study conducted by the team led by Koczkodaj, published in 2017 (Koczkodaj et al., 2017). In this study, the AUC-ROC algorithm was applied to reduce the Beck Depression Inventory (BDI) from 21 items to just 7, without compromising the test's validity or reliability. Moreover, the results indicated that the reduction in the number of items not only did not diminish measurement quality but actually improved it.

The AUC-ROC algorithm was used to select those items on the scale that showed the highest agreement with the external criterion, which in this case was a psychiatric diagnosis. Seven items were identified as the most diagnostic, having the greatest predictive value in relation to the external diagnosis. The remaining BDI items proved to be less diagnostically valuable and, in some cases, even detrimental to the overall score.

The use of the AUC-ROC algorithm enabled not only an effective reduction of the scale but also an increase in the validity and reliability of the measurement results. This approach was verified using the Graded Response Model (GRM) and Confirmatory Factor Analysis (CFA), both of which confirmed that the reduced scale did not suffer in quality. In fact, the validity and reliability of the measurement were found to have improved—an important achievement in the context of psychometric tool optimisation (Koczkodaj et al., 2017).

The results of this study, published by Koczkodaj and colleagues in 2017, represent a significant contribution to the field of psychometrics, demonstrating how mathematical algorithms can support the development of more efficient diagnostic tools.

Atypicality of responses is one of the major challenges in psychological measurement, potentially affecting the validity of the obtained results. Aranowska pointed out this issue, emphasising that atypical individuals—those whose traits manifest in ways that deviate from the classical patterns on which psychological tests are based—may be misjudged (Aranowska, 2016).

Psychological tests are generally designed to measure the typical, classical manifestation of a given construct or trait. However, problems arise when individuals in the population exhibit a trait in a somewhat different, atypical manner. In such cases, tests may fail to diagnose these individuals appropriately, resulting in reduced measurement validity. This highlights the need to develop more flexible diagnostic tools capable of identifying and accounting for such variations.

In such instances, artificial intelligence algorithms—especially decision tree algorithms—may prove helpful. These algorithms, by generating classification rules, can detect individuals who display a given trait in an atypical manner. Decision trees can analyse data in a way that enables the identification of patterns corresponding to atypical manifestations of traits. As a result, it becomes possible to identify individuals within a dataset who differ from the rest of the population in the way the trait is expressed.

This type of approach enables a more precise adaptation of diagnostic tools to the diversity of trait manifestations within a population, which can significantly enhance measurement validity. Accounting for response atypicality in psychometric analyses is essential for achieving more accurate and diagnostically meaningful results that reflect the full spectrum of how a given trait may present across individuals.

The chapter entitled *"Contemporary Approaches to the Construction of Psychological Scales Using Artificial Intelligence"* will discuss how artificial intelligence algorithms can support the scale construction process in ways that also accommodate atypical cases, thereby allowing for even more precise diagnosis of various psychological traits.

CHAPTER 36

Expert Systems as Intelligent Tests

The duration of psychological assessment has a direct impact on its quality and on the psychometric properties of the tools used. In particular, it affects the reliability and validity of tests, which are fundamental to their diagnostic value. Although the topic of psychometric test optimisation is widely discussed in the literature (Anastasi & Urbina, 1999; Aranowska, 2005), the use of expert systems in psychometrics remains an area requiring further research. In my earlier works, I have pointed to the potential of artificial intelligence and expert systems in psychological diagnosis and psychotherapy (Szymańska, 2024b), yet their application in the context of psychometric testing has not been comprehensively developed. This chapter aims to fill that gap and present the concept of intelligent tests based on expert systems.

Currently, one of the key challenges in psychometrics is finding a balance between test length and diagnostic validity. Overly lengthy scales, such as the MMPI-2, are examples of tests that may lead to distorted results due to respondent fatigue, which decreases both the accuracy of responses and the quality of psychometric data collected.

As test-takers proceed through lengthy assessments, they begin to experience fatigue, which leads to reduced attention, motivation, and response accuracy. In practice, this means that the respondent no longer carefully analyses subsequent test items but begins to respond in a biased, mechanical, or disorganised manner in an effort to complete the test as quickly as possible. This directly affects test reliability, as the results obtained in the latter sections of the test may not reflect the respondent's actual traits but rather the effects of fatigue and the desire to conclude the task.

The consequences of this mechanism are far-reaching. If a respondent begins to answer randomly, carelessly, or in a biased manner, the entire diagnostic process

is affected. Based on distorted data, it is impossible to make an accurate diagnosis, as the test ceases to measure what it was designed to measure. Its validity declines, since the results no longer reflect the respondent's true psychological state but rather the effects of fatigue and reluctance to continue.

As a result, a paradox emerges with long psychometric tests: instead of improving diagnostic quality through a greater number of items, they lead to worse outcomes, because respondents no longer answer in a manner consistent with their actual condition. This represents a fundamental issue with traditional diagnostic tools, one that calls for a solution in the form of modern, intelligent testing systems. For this reason, contemporary psychometrics increasingly aims to shorten even standardised scales to optimise assessment and reduce the burden placed on respondents (Koczkodaj et al., 2017).

While some psychological scales incorporate response quality control mechanisms in an attempt to minimise the issue of reduced reliability due to the length of assessment, limitations remain. A prominent example is the MMPI-2, which includes control scales designed to detect inconsistency in responses, chaotic answering patterns, or attempts at deliberate manipulation. These scales make it possible to identify situations in which the respondent did not provide coherent or trustworthy responses. However, a critical problem is associated with this approach: if the control scales detect a high level of response disorganisation, the result of the entire test is deemed uninterpretable (Kucharski & Gomula, 1998). The individual's profile is then rejected, as it fails to meet the reliability criteria required for psychometric analysis. This means that hours of testing may ultimately prove entirely useless, and the obtained data unsuitable for further interpretation.

This raises a fundamental question: what is the point of subjecting individuals to hours-long testing if there is a risk that the results will ultimately be discarded? Moreover, chaotic responses do not always stem from a conscious attempt to manipulate the test—they are often simply the result of fatigue, decreased concentration, and a natural reluctance to continue. This implies that the problem lies not only in the response style, but also in the measurement system itself, which fails to account for the limits of human attention and cognitive endurance.

This constitutes another argument in favour of re-evaluating the length of psychological tests. The goal is not to reject long scales altogether but to find a solution that enables more efficient measurement without the risk of complete data loss.

That solution lies in intelligent tests based on expert systems. Their implementation may allow for the dynamic adaptation of test length to the individual respondent, eliminating the need for long, static assessments. This chapter is dedicated to exploring precisely that issue.

The aim is not to shorten long psychometric scales, but rather to change the way we conduct assessment. Reducing test length would be equivalent to discarding the achievements of modern psychometrics, as in many cases, a comprehensive capture of a trait requires a broad set of test items. Reducing the number of questions could result in the loss of essential information, negatively impacting diagnostic

precision and the ability to identify substructures of the trait (Anastasi & Urbina, 1999; Aranowska, 2005).

Thus, the problem does not lie in the length of the scale per se, but in **the static nature of traditional psychometric tests.** In classical approaches, every respondent answers the same set of questions, regardless of their responses in the initial phase of the test. This is an inefficient method, as it fails to consider the individual course of assessment—some individuals end up answering too many redundant items, while others may not receive a sufficient number of relevant questions tailored to their profile.

Rather than abandoning long tests altogether, it is essential to develop a method that allows for flexible and precise adaptation of measurement to the individual respondent. Assessment must be conducted intelligently, in a way that accounts for both diagnostic needs and the individual's cognitive capabilities.

Contemporary psychometric tests are structurally static. The respondent begins the test and answers all questions in a predetermined order, regardless of how their results unfold during the assessment. The psychologist receives the full set of responses only after the test is completed and can analyse the results only at that stage.

Yet in many cases, it is evident that continuing the assessment in its entirety is unnecessary. If a scale contains 10 items and the respondent scores at the lowest level on the first five, then even if they respond at the highest level to the remaining five, their final score will not change substantially enough to alter the diagnostic interpretation. They will not suddenly achieve a high score that shifts the overall result—their outcome will remain low or moderate. In such cases, an intelligent measurement system would have no reason to present additional questions that add nothing to the diagnostic interpretation. It would be sufficient to terminate the assessment at that point and, based on the available data, assign the result to the appropriate diagnostic category.

Such a mechanism would not only reduce the duration of the assessment but also alleviate unnecessary cognitive load on the respondent. In a scale with a large number of items, this could reduce the number of questions by as much as half while still preserving the full diagnostic value of the test. As a result, testing would become more efficient and aligned with the actual needs of the respondent, eliminating redundant questions without compromising the quality of the results. An intelligent test would present subsequent items dynamically and interrupt the assessment once analysis of the existing responses indicated that the outcome was already situated in the low or moderate range, with no potential for a meaningful change in diagnostic interpretation.

In addition to eliminating unnecessary questions, intelligent tests could also dynamically deepen the assessment in areas that prove diagnostically relevant. If the initial responses indicate that a respondent is scoring within the depression range, the system could automatically introduce follow-up items to more precisely verify the severity of symptoms. This would enable a comprehensive and accurate

diagnosis without wasting the respondent's time or requiring subsequent clarification of results. It would also eliminate the need to invite the respondent back for further testing—the test would respond immediately to elevated scores by automatically extending measurement in key areas.

Such a mechanism enables the diagnostic tool to be precisely tailored to the individual results of the respondent. In traditional tests, the respondent proceeds through all questions, even if some are redundant or the answers are already evident. In contrast, here the test not only reduces questions in irrelevant areas but also expands the sections of the assessment that prove to be diagnostically significant.

For example, if a respondent initially scores high on a depression scale, the test may introduce differentiating items to determine whether this is a random deviation or a clinically meaningful issue. Strengthening the measurement allows for more accurate diagnosis and reduces the risk of false positives.

A similar mechanism can be applied in other scales. If the initial items on a schizophrenia scale indicate that the respondent does not meet diagnostic criteria, the test may automatically skip further items in that domain. However, if the responses suggest the possibility of symptoms, the test can add supplementary questions to refine the diagnosis.

In this way, testing becomes not only more efficient but also more precise, as each respondent receives a set of questions tailored to their individual diagnostic profile as it emerges during the assessment.

Achieving an intelligent test that continuously analyses the respondent's answers, adjusts questions, and processes results in real time is possible through the use of existing expert system methodologies. Expert systems are advanced computational structures (interactive knowledge bases) that emulate the decision-making processes of experts in a given field (Rutkowski, 2006). Their architecture can serve as the foundation for intelligent psychometric tests, enabling the dynamic adaptation of the diagnostic process.

An expert system consists of several fundamental components:

1. **User interface** – the module that enables communication between the respondent and the system. In the context of an intelligent psychometric test, this would be the interface through which the respondent answers subsequent questions.
2. **Knowledge base** – stores the information and decision rules upon which the system operates. In the case of intelligent tests, the knowledge base would contain question structures, diagnostic criteria, and the relationships between responses and the further progression of the test.
3. **Inference engine** – the key module of the system responsible for analysing responses and making decisions regarding the subsequent course of the test. It may utilise logical rules, such as *modus ponens* (if A, then B) and *modus tollens* (if not B, then not A), as well as probabilistic algorithms that dynamically adjust the test based on the respondent's prior answers. The combination of these methods allows for intelligent adaptation of the testing process—eliminating unnecessary questions and reinforcing diagnostically relevant areas.

4. **Explanation module** – an optional component of the system that can provide information on the basis for a particular diagnostic decision or explain why certain questions were omitted.

In the context of an intelligent test, the expert system would operate as follows:
1. The respondent begins the test, and the system presents the initial questions.
2. The inference engine analyses the responses and, based on them, determines the subsequent questions:
   - If the responses in a given scale indicate a low score, further questions may be omitted.
   - If the responses suggest a potential diagnostic outcome, the system may add supplementary items to increase diagnostic precision.
3. Decisions are made in real time based on the rules contained in the knowledge base and the applied inference methods.
4. Upon completion of the test, the system automatically generates an interpretation of the results, taking into account both the responses given and the way the test was dynamically adapted during the process.

Why is the expert system an ideal framework for intelligent tests? Expert systems are already in use in fields such as medicine, technical diagnostics, and decision automation across various domains (Michalik, 2006b). Their application in psychometrics would allow for the development of tests that are not only flexible but also diagnostically more precise. Thanks to inference mechanisms, the system can detect subtle relationships between responses and automatically adjust the course of the test to maximise its efficiency.

In summary, intelligent tests could employ the architecture of expert systems to dynamically adapt the diagnostic process in real time. This methodology enables the optimisation of test length, the elimination of unnecessary items, and the reinforcement of diagnostically significant components—leading to more reliable and valid results.

In her 2024 chapter, Szymańska highlights the key role of expert systems in psychometric diagnostics, emphasising their ability to optimise the testing process (Szymańska, 2024b). She argues that the use of expert systems in psychology may significantly improve both diagnostic validity and the efficiency of psychometric research. However, it is important to note that expert systems are not limited to rules provided by researchers—they also possess the capacity to independently generate and modify rules through artificial intelligence algorithms.

Expert systems are based on two fundamental types of decision rules:
1. **Static rules** – provided by the knowledge engineer during the construction of the system (Michalik, 2006b). These are principles grounded in prior research and psychometric theory, which define how the test should function, which questions should be asked, and what relationships exist between variables.
2. **Dynamic rules** – generated during the system's operation by artificial intelligence

algorithms (Michalik, 2006b). Through the application of decision tree algorithms (inductive algorithms) and predictive algorithms such as artificial neural networks, the system can continuously analyse collected data and formulate new diagnostic rules that were not originally anticipated. Thanks to these algorithms, tests can operate not only on the basis of predefined rules but also learn and adapt as additional data are gathered.

The use of artificial intelligence in expert systems thus enables an even more comprehensive psychometric analysis, combining the classical diagnostic approach with modern data analysis methods. This integration allows for the dynamic adaptation of psychometric tools to the individual needs of the respondent, ensuring greater diagnostic validity and precision.

In summary, the concept of expert systems as intelligent psychometric tests presented in this chapter opens new perspectives in psychological assessment. Their ability to dynamically adjust the testing process, eliminate unnecessary items, and enhance diagnostically relevant areas represents a significant step toward more effective and valid psychometrics. However, as with many of the applications of artificial intelligence in psychometrics discussed in this part of the book, the implementation of such solutions requires further research and testing.

Key questions remain regarding the validation of these new methods, their reliability, and their impact on diagnostic standards. Further work is also needed on integrating artificial intelligence with psychometric test theory to ensure not only the automation of measurement but also its alignment with scientific and practical requirements.

CHAPTER 37

Artificial Intelligence in Profiling Psychological Traits: New Approaches in Diagnostics and Psychometric Modeling

Psychological profiling has long been an important tool in criminology, forensic psychology, and behavioural analysis. However, its methodology remains more intuitive than mathematical. As noted by Łozińska-Piekarska & Dąbrowski (2023), profiling raises concerns regarding its effectiveness and scientific foundations. It largely relies on subjective interpretation of data and the experience of profilers, which limits its evidential value in legal proceedings. The process involves constructing a psychological profile of an individual—most commonly an offender—based on the analysis of behaviours, modus operandi, and criminal patterns. Profilers in the FBI, police, and forensic teams analyse behavioural traces, that is, recurring traits and habits, through which they attempt to infer personality characteristics, motivations, and potential future actions of the individual.

A major issue in classical profiling, however, is the absence of clear, measurable principles for generating profiles. The process is primarily based on expert knowledge, heuristics, and the profiler's experience, which makes it inherently subjective and susceptible to cognitive biases. As Łozińska-Piekarska & Dąbrowski (2023) point out, profilers often reach different conclusions depending on their backgrounds and adopted methodologies. This stems not only from the lack of formal profiling standards but also from the absence of unified educational frameworks for future profilers, both of which hinder the reliability and reproducibility of the process. Although profilers make use of statistical data and the analysis of previous cases, there are still no clearly defined mathematical models capable of generating psychological profiles in an objective and repeatable manner based on data. At present, profiling

continues to rely heavily on intuition and the individual interpretation of available information.

As a result of this approach, profilers may arrive at divergent conclusions depending on their experience and methodology, thereby impeding the standardisation of the process. Moreover, the lack of formalised models makes it difficult to verify the accuracy and effectiveness of profiling. Łozińska-Piekarska & Dąbrowski (2023) emphasise that the resulting profiles are general in nature and cannot be treated as tools for definitive offender identification. In Poland, profiling is regarded as an auxiliary method and does not hold the status of hard evidence in legal proceedings, which contributes to scepticism among law enforcement authorities regarding its efficacy. There are no clear instruments for assessing the extent to which a profile actually fits the offender or its real impact on the effectiveness of investigations.

A more objective approach to profiling is therefore a key concern in psychometric analysis and reliable measurement. The central challenge is to develop a profiling model that does not rely solely on intuition and expert judgement, but rather on objective data analysis methods that allow for replicable results. While traditional criminal profiling raises concerns regarding its validity and effectiveness (Łozińska-Piekarska & Dąbrowski, 2023), the application of artificial intelligence could facilitate the objective analysis of behavioural patterns and their mathematical modeling.

To apply artificial intelligence in profiling, the first step is to gather appropriate data. It is crucial to distinguish between population-level data and data specific to, for instance, offenders. Criminal profiles cannot be constructed based on general psychological data, as offenders do not constitute a representative group of the broader population. To create valid profiling models, one must draw directly from sources that contain information about individuals who have actually committed crimes.

Such data are found in archives, investigative records, court files, criminal biographies, psychological reports, forensic literature, and crime descriptions. These materials are unstructured, meaning they are not presented in a form ready for analysis but are dispersed across various texts, reports, narratives, and databases with varying levels of formalisation.

Only after gathering these materials is it possible to construct large datasets—so-called Big Data—that form the foundation for further analysis (Stephenson, 2018). Achieving this requires the use of algorithms capable of processing unstructured data and transforming it into structured form (Elder et al., 2012; Nisbet et al., 2009; Szymańska, 2017b; Szymańska & Aranowska, 2019). This process involves natural language processing (NLP), the extraction of key information, and the systematic classification of data in a format suitable for subsequent modeling.

An essential stage in converting unstructured into structured data is the initial text preprocessing using language models. Before the data are encoded numerically, NLP algorithms can be applied to automatically analyse texts and extract key psychological and behavioural features.

Language models such as BERT or GPT may be used to extract relevant information from documents, court records, crime descriptions, and offender biographies.

Their function is not only to process the text but also to identify linguistic and structural patterns that may indicate specific personality traits, motivations, or behavioural tendencies of offenders.

Through this process, language models enable the automatic categorisation of individuals and the assignment of numerical values to textual descriptions, which can then be used in statistical analysis. For example, the system may analyse offender descriptions in terms of keywords related to impulsivity, manipulation, or lack of empathy, thereby allowing for preliminary psychological classification of a given individual.

The reader of this book is already familiar with the fundamentals of algorithms used in text mining, so it can now be noted that language models serve as a preliminary filter in this process. They enable the extraction of qualitative features and their transformation into a form ready for numerical analysis. After this preprocessing stage, the data can be recorded in a structured database, making them suitable for processing with advanced artificial intelligence methods and statistical techniques.

In this way, AI-based profiling gains a solid foundation—rather than relying solely on intuition, it is grounded in hard data acquired and processed through language algorithms and text analysis.

By transforming unstructured data into a structured database, we can proceed to the next stage—actual profiling. This database now serves as the foundation for deriving model profile curves that describe specific disorders, psychopathic traits, or other personality patterns.

Mathematical profiling differs from the classical intuitive approach in that it is based on empirical data and clustering algorithms, which allow for the identification of natural groupings among analysed cases. One of the primary tools is the *k-means* method, which partitions data into *k* optimal clusters based on similarity. Alternatively, the *Expectation-Maximization (EM)* algorithm can be used, offering more flexible modeling of relationships and probabilistic classification of individuals.

Thanks to these algorithms—described in previous chapters of this book—it becomes possible to derive model curves (profiles) corresponding to specific personality types or disorders. For example, one can construct a psychopathy profile based on characteristic features such as lack of empathy, impulsivity, or low emotional regulation. The system may then compare newly analysed cases with existing model curves and assign individuals to specific profiles with a corresponding degree of probability. This process involves matching the empirical curve derived from the new case with the relevant model curve, enabling the objective classification of individuals based on their behavioural and psychological traits (Aranowska, 1989; Szymańska, 2023b).

This entire process has nothing in common with intuitive analysis—it is a strictly mathematical approach, fully driven by artificial intelligence algorithms. At every stage of analysis—from initial text processing to classification and profile modeling—decisions are made by algorithms, not by subjective interpretation.

As a result, AI-based profiling not only provides a tool for detecting behavioural patterns but also offers an objective and measurable method of analysis, in which every decision stems from a precise mathematical algorithm rather than loose data interpretation.

In summary, the application of artificial intelligence in profiling opens a new chapter in diagnostics and behavioural analysis. Instead of relying on experts' intuitive inferences based on subjective experience, it becomes possible to create an analytical system that operates from start to finish on mathematical principles. As noted by Łozińska-Piekarska & Dąbrowski (2023), current profiling methods are largely intuition-based and lack clearly defined standards. The introduction of artificial intelligence methods may be key to increasing the reliability and reproducibility of psychological profiling.

Each stage—from transforming raw textual data into a structured database, through the use of language models to extract features, to identifying patterns in the criminal population using clustering algorithms—is conducted in an objective, transparent, and replicable manner. This means that profiling ceases to be a "black box" in which classification decisions are difficult to verify, and becomes a precise diagnostic tool with clearly explainable and verifiable foundations.

This approach frames profiling as a process that can be mathematically modelled, tested, and developed. Artificial intelligence introduces a new quality—making profiling not only more efficient, but above all scientifically grounded and replicable. As such, it redefines profiling as a systematic process of analysis that can be applied in diagnostics, criminology, and psychometrics, yielding scientifically supported, objective results.

It is worth noting that the first attempts to apply artificial intelligence to the mathematical profiling of personality types have already been initiated. One example is the master's thesis by Maja Szczygieł (2024), conducted under my supervision, in which a model was developed to classify fifteen disordered personality types based on diagnostic descriptions of serial killers. The input data consisted of detailed qualitative case descriptions, which were subsequently processed using language models and classification algorithms. The results confirmed that the proposed method enables accurate and consistent identification of an individual's personality profile—even in such a complex domain as personality disorders among perpetrators of the most serious crimes. This methodology thus validates the approach presented in this chapter, demonstrating that AI-based profiling can indeed be implemented in practice, delivering measurable, reproducible, and scientifically justified results.

CHAPTER 38

Advanced Issues in Psychometrics Supported by Artificial Intelligence

Contemporary psychometrics faces the challenge of advancing to a higher level of precision in the description of psychological phenomena. Traditional approaches have relied on two-dimensional analyses of psychological traits, stemming both from methodological limitations and from historical assumptions underlying the field of psychometrics itself. This chapter explores the application of topological approaches in psychometrics and the analysis of psychological measurement in three-dimensional space. Introducing this perspective allows for a more accurate representation of the structure of psychological traits and their interrelationships—relationships that, in classical models, have often been examined solely within two-dimensional frameworks.

To understand the groundbreaking nature of the three-dimensional approach, it is first necessary to examine how the concept of dimensionality has been understood in psychology thus far. Classical psychometric models conceptualised psychological traits as one-dimensional continua—placing the individual along an axis stretching between two extreme values. A typical example is the introversion–extraversion dimension, where the level of extraversion is defined as higher or lower along a single axis. While such an approach may be useful for basic diagnostic purposes, it fails to reflect the complexity of psychological functioning, which is multifactorial and dynamic.

Notably, Carl Gustav Jung did not view introversion and extraversion as polar opposites on a single scale, but rather as qualitatively distinct attitudes toward the world, co-occurring with other aspects of personality. In this framework, the individual is not "on an axis" but in a configuration of traits. Similarly, Herbert Shuey emphasised that Jung's types are phenomenological and cannot be represented as

a continuum, as they lack a structural biological basis (Geyer, 2012). This understanding opens the door to topological approaches that enable the modeling of trait relationships within spaces of greater dimensionality than those allowed by traditional two-dimensional analysis.

Multidimensional personality models, such as the Five-Factor Model, might at first glance appear to overcome this limitation, as they assume the existence of several independent personality dimensions. However, in psychology, the concept of multidimensionality is understood differently than in mathematics. When psychometrics refers to multiple dimensions, it typically denotes distinct traits measured in parallel—each on its own continuum. This means that although the model encompasses several traits simultaneously, each still exists along a separate axis, and the relationships between them are analysed primarily through correlations.

The mathematical concept of dimensionality refers to the number of independent directions in a space. In two-dimensional space, any point can be described using two coordinates, with each variable corresponding to one of the coordinate axes. Introducing a third dimension means adding a new axis that is linearly independent from the others. This shift moves the analysis from a flat plane to a space in which each point is defined by three coordinates. As a result, relationships among variables can be described not only as distances on a plane but also as dependencies within a structure of greater dimensionality—allowing for the capture of more complex patterns (Szymańska, 2025e).

This very distinction between the psychological and mathematical understanding of dimensionality is crucial for grasping the proposed shift, which demonstrates why topological tools may more accurately reflect the relationships between traits in psychometrics. Traditional models describe relationships between variables within a two-dimensional space, whereas topological approaches allow for a more precise representation of complex dependency structures—particularly in the case of traits that cannot be adequately captured in spaces with too few dimensions. In certain instances, transitioning to three-dimensional or higher-dimensional spaces becomes necessary to accurately model the actual relationships between variables—relationships that may otherwise be flattened into oversimplified linear associations in classical analyses.

The aim of this chapter is to present specific cases in psychology in which a three-dimensional perspective becomes a necessity. Where traditional approaches have failed to capture complex interrelations between traits, new methods make it possible to develop more adequate models that account for the true dynamics of psychological processes. In the following sections, we will examine concrete examples and the mathematical tools that enable such an approach, thereby opening a new chapter in psychometrics.

Until now, the prevailing approach in psychometrics has assumed the analysis of trait space within a two-dimensional framework, which has carried specific methodological consequences. The verification of such spaces has typically been conducted using factor analysis, a method that allows for dimensionality reduction

of psychological data by identifying latent structures underlying observed variables (Aranowska, 2005). In classical factor analytic models, the goal is to identify the main axes of variance in the dataset, thus isolating the fundamental psychological dimensions—such as the personality factors of the Five-Factor Model.

However, with the advancement of psychometrics and the increasing complexity of cognitive models, there emerges a growing need to account for three-dimensional spaces—where classical tools like factor analysis are no longer sufficient (Szymańska, 2025a). This calls for a new mathematical approach capable of classifying data in higher-dimensional spaces and describing relationships that cannot be captured within two-dimensional frameworks. In such cases, traditional factor analysis proves overly simplistic, as it assumes that relationships between variables can be reduced to simple linear equations within two dimensions (Szymańska, 2025c).

The development of methods that allow for psychological modeling in higher-dimensional spaces will be closely tied to advances in artificial intelligence. AI provides tools for analysing large datasets and detecting nonlinear patterns that may reveal dependencies invisible to classical statistical methods.

38.1. The Limits of Classical Psychometrics: Artificial Intelligence and Multidimensional Models

Contemporary psychometric models that describe the structure of multidimensional traits encounter significant limitations in verification—particularly in the case of circular (circumplex) models. Although such structures theoretically enable precise representation of inter-variable relationships, they become analytically problematic as the number of dimensions increases and the variables are arranged at sharp angles to one another in space (Szymańska, 2025a).

One of the core challenges arises when dimensions are **tightly compressed**—that is, when correlations between variables are high and the angles between their corresponding vectors are small. In such cases, empirical separation of the variables becomes increasingly difficult. The cosine of the angle alpha, which corresponds to the correlation coefficient, reaches higher values as the angle between variables decreases. This results in challenges in reliably distinguishing independent traits. Consequently, classical psychometric methods may prove inadequate for properly verifying such complex structures, as emphasized by Szymańska (2025a, 2025c) in her study.

This issue calls for a new methodological approach capable of navigating multidimensional spaces where traditional techniques reach their limits. A central challenge in such models is the selection of a statistical framework that allows for precise discrimination between highly correlated variables—e.g., those correlating at 0.90 or 0.92. In classical psychometrics, such strongly correlated variables are typically assumed to measure the same construct. However, even minor angular differences between them may indicate meaningful conceptual distinctions.

To better illustrate this issue, let us consider three traits—A, B, and C—arranged in a multidimensional space within a circular model. If variables A and B form an angle of 12 degrees, this implies a correlation of approximately 0.978, since the cosine of 12° is $cos(12°) \approx 0.978$. Meanwhile, if variables A and C are separated by an angle of 17 degrees, their correlation drops to around 0.956 ($cos(17°) \approx 0.956$). From the standpoint of classical factor analysis, the difference between these two correlations—0.978 and 0.956—may appear negligible. However, in the context of multidimensional models, such angular differences translate into significant psychometric distinctions.

It is precisely in such cases that classical psychometrics reaches its limits, as it primarily operates within the frameworks of factor analysis or regression—both of which assume linear relationships between variables. However, analysis in spaces with higher dimensionality requires methods that account not only for the strength of correlations but also for their topological configuration.

In the case of a circular model involving 15 variables, each variable must be positioned at equal angular distances from its neighboring variables. This implies that the angle between adjacent variables equals 24 degrees. When converted into a cosine, this yields a value of approximately 0.914—indicating that the variables are correlated at the level of 0.91.

Even though such a model can theoretically be constructed, verifying its correctness within a two-dimensional framework becomes virtually impossible using classical psychometric methods. The problem is no longer limited to distinguishing variable A from B, but also includes accurately differentiating between A and C, A and D, and so forth, in accordance with the entire model structure. All variables must be distributed in alignment with the assumed correlation values, and as the number of variables increases, achieving this becomes methodologically burdensome.

This difficulty arises from the fact that classical verification techniques are unable to reliably distinguish between variables with such high intercorrelations. Their discriminability decreases as the number of variables in the model increases. Consequently, classical psychometrics does not offer adequate tools for the empirical analysis of such structures, necessitating the search for alternative research methodologies.

Szymańska (2025a) addresses this issue in her article, highlighting the fundamental limitations of classical psychometrics in the validation of circular models. She points out that when variables are arranged at very narrow angles and their correlations exceed 0.91–0.95, traditional statistical methods cannot effectively differentiate between them. In response, Szymańska proposes the application of artificial intelligence algorithms that allow data to be projected into higher-dimensional spaces, thereby improving the separation of variables and overcoming the issue of dimensional compression.

One proposed solution in this context is the use of Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel (Szymańska, 2025c). The

application of a nonlinear kernel enables the projection of data into a higher-dimensional space, in which variables that were initially difficult to distinguish can be more effectively separated. As a result, the multidimensional model gains in precision, and the classical constraints of two-dimensional verification cease to be a limiting factor.

Relaxing such highly correlated variables through appropriate space transformation enables a more realistic representation of the structure of psychological traits. As a result, not only does the effectiveness of classification and model validation improve, but new possibilities also emerge in the analysis of multidimensional psychometric configurations.

However, to effectively employ the method of projecting variables into higher-dimensional space, it is necessary to revisit the psychometric models themselves in terms of their actual three-dimensional or multidimensional nature. If we assume that applying the RBF kernel (or other nonlinear transformations) improves the separation of variables in three-dimensional space, the key question remains: is this shift into a higher dimension theoretically justified, or does it merely serve as a mathematical analytical tool?

If our psychometric models do not account for the actual three-dimensional structure of psychological traits, then such a mathematical procedure—though effective in separating variables—adds no theoretical value. The algorithm may enhance the quality of the analysis, but in itself it does not explain why psychological traits should be organized in higher-dimensional space. In other words, one may achieve better empirical classification, but without a theoretical foundation, this remains a purely technical statistical solution, not a genuine reflection of the underlying structure of the traits being studied.

Therefore, to fully leverage the potential of artificial intelligence in psychometrics, it is necessary to simultaneously develop theoretical models that justify this approach. If it can be demonstrated that the structure of psychological traits is indeed three-dimensional and requires a higher-dimensional framework for full analysis, then the use of methods such as SVM with an RBF kernel becomes not only a mathematical tool but also a method aligned with the actual nature of psychological processes. Otherwise, AI in psychometrics will remain merely a way to improve variable separation in models, without contributing to a deeper understanding of their structure.

The advancement of psychometrics toward multidimensional models cannot therefore rely solely on artificial intelligence tools—it also requires a redefinition of theoretical foundations that clarify the conditions under which a three--dimensional approach is justified and what implications it holds for psychology as a science.

38.2. Psychometric Models in Higher Dimensions: How Artificial Intelligence Can Expand the Measurement Space

Contemporary psychometrics continues to operate within models that assume a two-dimensional or multidimensional, yet still linear, structure of traits. However, as measurement methods and psychological theories evolve, it is becoming increasingly evident that many mental processes and theoretical constructs may require a three-dimensional—or even more complex—representation. Once psychology defines or discovers models that truly incorporate higher-dimensional spatial structures of traits, there will arise a need for tools capable of verifying these models and translating them into actual measurements.

At this stage, artificial intelligence may play a pivotal role. Not only does AI allow for the analysis of nonlinear relationships between variables, but it also enables the expansion of the measurement space through algorithms capable of operating in higher dimensions. Classical statistical approaches may prove insufficient for capturing relationships that are nonlinear, multidimensional, and dynamic; therefore, it becomes necessary to employ methods that allow for more flexible modeling of the trait space.

This section focuses on which AI algorithms can be applied in psychometrics to enable operation within expanded dimensional spaces, and how such technologies help overcome the limitations of classical psychometric approaches. The use of methods such as support vector machines, neural networks, and deep learning algorithms may open new avenues for analyzing and interpreting psychological traits in ways that were previously inaccessible.

One such algorithm that enables the expansion of the measurement space in psychometrics is the support vector machine (SVM), with its key mechanism—the Radial Basis Function (RBF) kernel (Hearst et al., 1998). The RBF kernel serves a crucial function: it allows data originally embedded in a lower-dimensional space (e.g., 2D) to be projected into a higher-dimensional space (e.g., 3D), where it becomes easier to separate individual variables.

The operation of this algorithm relies on the fact that, in the original (input) space, data points may lie very close to one another and be difficult to distinguish. In classical psychometrics, where variables are analyzed in two-dimensional systems, there is often a problem of separating traits with very high correlations. The RBF kernel addresses this issue through nonlinear transformation, projecting the data into a higher-dimensional space where traits that were previously too closely aligned in lower dimensions can now be effectively separated.

A key element in this process is the *separating hyperplane*—the boundary that separates different classes of data once they have been mapped into a higher-dimensional space (Hearst et al., 1998). In the classical SVM algorithm operating within a linear space, the hyperplane is simply a line separating two classes in two

dimensions or a plane separating them in three. However, in more complex cases where the boundaries between variables are nonlinear, the RBF kernel transforms the data in such a way that separation becomes feasible in a higher-dimensional space.

This process can be explained through the concept of *support vectors*. Support vectors are the data points closest to the separating hyperplane and are the ones that determine its position. In classical SVM, the algorithm maximizes the margin between these support vectors and the hyperplane, enabling more precise separation of variables in the space.

In the context of psychometrics, this means that if psychological traits are too closely positioned in a two-dimensional space, their separation using classical methods becomes difficult or even impossible. The RBF kernel transforms such data by projecting them into three dimensions, where separation becomes feasible through a newly defined hyperplane in the higher-dimensional space. This makes it possible to distinguish traits that were previously nearly indistinguishable in a lower-dimensional configuration.

The practical application of this approach opens entirely new possibilities in psychometric analysis, allowing for more precise modeling of traits and their interrelations within an extended measurement space.

Such a solution enables the precise separation of variables, revealing their actual position in space. As a result, it becomes possible to unfold previously compressed dimensions, allowing for a more accurate analysis of the model's structure. Szymańska (2025a) applied this approach in her study on Gurycka's theory of parental mistakes, whose model is based on a circular structure but in fact has a three--dimensional character (Szymańska, 2025d).

This model was originally described by Gurycka as a two-dimensional (circular) structure; however, in her analysis, Szymańska demonstrated that its actual organisation requires a three-dimensional representation, which justifies the application of an RBF kernel for its verification. A key element of this structure is the dichotomy running along the radius of the circle – parental mistakes are described on the outer part of the circle, whereas positive parental attitudes are located inside the circle (Gurycka, 1990). This means that the variables in this model are not arranged on a single plane but form a hierarchy, in which the values closer to the centre carry a different interpretation than those located at the periphery.

For this reason, classical statistical approaches encounter difficulties in accurately classifying variables, as observations located near the centre of the circle are characterised by low variance and may be treated as diagnostically irrelevant. Aware of the three-dimensional structure of the analysed model, Szymańska proposed the application of a radial basis function (RBF) transformation, which enables the mapping of data into a higher-dimensional space. By using a Gaussian kernel, it becomes possible to project these data into a Hilbert space, allowing the capture of nonlinear relationships between variables. In such a constructed space, features with opposite poles—previously positioned on a single plane—become geometrically separable in a precise manner. This, in turn, allows for a more accurate analysis of their

mutual relationships, improves class separability, and enables the identification of hidden patterns that could remain undetected within a classical spatial framework (Szymańska, 2025a).

As a result, Gurycka's model of parental mistakes, instead of being analysed as a two-dimensional map of dependencies, could be transformed into a structure accounting for the hierarchical and spatial relations between parental attitudes (both erroneous and appropriate). This approach not only allowed for a better interpretation of the results, but also demonstrated that many psychological models considered two-dimensional may, in fact, require description within a three-dimensional space (Szymańska, 2025d).

One such model that most likely also possesses a three-dimensional structure is Olson's model, which is based on assumptions similar to those in Gurycka's framework. In its structure, traits located inside the circle are qualitatively different from those placed on its outer part (Lachowska, 2008). It can therefore be inferred that their actual interrelations may also require analysis in three-dimensional space (Szymańska, 2025d).

Factor analysis, as one of the fundamental statistical methods used in psychometrics, assumes that the origin of the coordinate system—that is, zero—represents an absence of variability or its minimal intensity. This means that traits measured within such a model are treated as a continuum: the farther from zero on the coordinate axes (X and Y), the more developed the trait is considered. Values located near zero are interpreted as low-intensity traits, whereas those more distant from the centre are regarded as strongly developed psychological properties.

Such a structure makes factor analysis effective for describing variables in a two-dimensional space, but only when the examined traits indeed align along a one-dimensional continuum. The problem arises in the case of models in which the variables at the centre of the system are not weaker intensities of peripheral traits, but represent qualitatively different categories of variables.

Such a situation occurs in Gurycka's model, where positive parental attitudes are located in the inner part of the circle, and parental mistakes are positioned on the outer part. However, these are not values of differing intensity of the same trait, but rather two entirely distinct psychological constructs, which cannot be described as a continuum. Factor analysis, in attempting to encompass the entire model, would treat values near the centre as underdeveloped versions of those at the periphery, leading to a misinterpretation of the trait structure.

As a result, it is possible to statistically confirm either the structure of the variables located on the outer part of the circle or the structure of the variables within the circle—but not both simultaneously. Classical statistical approaches are unable to capture this type of relationship, as they operate within the framework of two-dimensional, linear dependencies between variables. Consequently, they are incapable of verifying models that assume variables positioned closer to the centre of the circle are qualitatively different, rather than merely less intense versions of the variables on the periphery.

For this reason, it becomes necessary to employ methods that are not limited to classical statistical analyses but allow for the capture of the multidimensional structure of the data, including their topology and spatial relationships.

What has been presented in this section constitutes only a preliminary outline of the problem, which requires further elaboration both at the theoretical and methodological levels. The challenges described here, related to the analysis of three-dimensional models such as Gurycka's, indicate the need not only to transfer measurement into higher dimensions but also to more precisely define the shape of that space.

Since these models require representation in three-dimensional terms, the natural next step is to determine their structure, identify appropriate methods for their visualisation, and propose ways of empirically verifying them. Questions concerning the types of relationships that govern this system, its actual shape within the psychometric space, and how it can be translated into specific analytical methods remain open and constitute a direction for future research.

CHAPTER 39

Contemporary Approaches to the Construction of Psychological Scales Using Artificial Intelligence

The construction of psychological scales has always been grounded in certain theoretical assumptions that serve as the foundation for developing measurement instruments (Anastasi & Urbina, 1999; Hornowska, 2003). However, a problem arises when there is no solid theory describing the given psychological construct. In such cases, it becomes necessary to thoroughly understand the phenomenon to be measured in order to construct a measurement tool that faithfully reflects its multidimensionality.

In this context, artificial intelligence (AI) proves to be extremely helpful. With a wide array of algorithms, ranging from speech recognition algorithms to inductive ones such as decision trees, AI enables not only the analysis of large data sets but also the extraction of meaningful patterns and the creation of rules that may describe a given psychological construct (Szymańska, 2017b).

In situations where theoretical knowledge is limited or incomplete, artificial intelligence (AI) may serve as a key tool in the process of discovering and describing psychological phenomena. By analysing vast amounts of data from various sources, AI is capable of identifying multidimensional aspects of a phenomenon that may be difficult to capture using traditional research approaches. This enables the creation of more comprehensive and accurate psychological scales that better reflect the reality of the constructs being studied (Rzechowska & Szymańska, 2017). In fact, AI can independently carry out a process similar to grounded theory.

Grounded theory is a research approach that involves generating theory based on the systematic analysis of empirical data, rather than testing pre-existing hypotheses. In traditional research, this process is carried out by researchers who iteratively analyse data, searching for patterns, categories, and relationships from which a theory gradually emerges (Babbie, 2007). Using AI to perform a similar process allows for the automation and significant acceleration of such research, as AI algorithms can analyse data on a scale that exceeds human capabilities while simultaneously identifying hidden patterns and dependencies that may lead to the discovery of new theories.

By fulfilling an analytical function, AI is capable of processing both qualitative and quantitative data, which enables the emergence of new hypotheses and theories in a more dynamic manner. This process can be regarded as a form of automated grounded theory discovery, in which AI, based on data, creates new theoretical models while enabling researchers to iteratively test and refine those models using subsequent data sets. This allows for a more flexible and innovative approach to psychological research, especially in areas where theoretical knowledge is still in its formative phase or where phenomena are too complex to be fully understood using traditional research methods (Szymańska, 2024b).

Moreover, AI is also extremely helpful in identifying and accounting for atypical manifestations of psychological traits. Thanks to its advanced analytical capabilities, AI can detect cases in which a given trait presents itself in an untypical manner—something that is often overlooked by traditional psychological tests (Aranowska, 2016). Taking such atypicalities into account in the construction of a scale enables the development of diagnostic tools that are more precise and better suited to the actual variability in how traits manifest across the population.

In 2004, Rzechowska published a breakthrough method known as the Reconstruction of Process Transformation, in which she applied an artificial intelligence algorithm—specifically, the C4.5 algorithm developed by Quinlan, which is an inductive decision tree algorithm (Rzechowska, 2004). This method was used to describe the pathways within a model that illustrate the multidimensionality of psychological processes. The C4.5 algorithm enabled the structuring of these pathways, presenting different variants of the course of a given process depending on the number of branches in the decision tree. Each branch of the tree corresponded to different scenarios and variants of the process, allowing for a deep understanding of the complexity and dynamics of the phenomena under study.

An example of the application of this method was a study conducted by Rzechowska on a group of individuals aged 50+, teenage mothers, and other populations, in which the C4.5 algorithm made it possible to identify various pathways of psychological processes within these groups (Rzechowska, 2002, 2011a, 2011b, 2011c). This method not only accounted for dimensional diversity but also allowed for the tracking and analysis of various possible trajectories of these processes. In other words, this approach made it possible to understand how one variant of a process could transition into another, and what the potential scenarios for such transitions might be.

This innovative method demonstrated how artificial intelligence—and decision tree algorithms in particular—can be useful in psychological research for describing and understanding complex mental processes. Thanks to the analytical power of AI, researchers were able to gain a more detailed and dynamic picture of the phenomena under investigation, which in turn opened new perspectives for both the construction of psychological scales and the analysis of psychological processes.

In 2017, Rzechowska, in collaboration with Szymańska, undertook a pioneering attempt to describe the construction of a segment of a psychological scale based on the previously mentioned method of the Reconstruction of Process Transformation (Rzechowska & Szymańska, 2017). In this endeavour, decision trees based on Quinlan's C4.5 algorithm played a key role in identifying different variants of experiences among teenage mothers.

In this process, the C4.5 algorithm identified attributes that represented various pathways and variants in the course of psychological processes. These attributes, indicated by the algorithm, were subsequently transformed into specific items in a psychological test. In other words, each branch of the decision tree, representing a specific attribute or trait, was directly used as the basis for creating questions within the psychological scale (Rzechowska & Szymańska, 2017).

In their work, Rzechowska and Szymańska demonstrated how advanced decision tree algorithms can be applied to the construction of psychological scales, enabling the development of more precise and tailored diagnostic tools. This method made it possible not only to understand the complexity of experiences within the studied groups but also to use this information to construct a scale that reflected those specific experiences in a valid and reliable way.

In this manner, the Reconstruction of Process Transformation strategy provided not only a tool for analysing the course of psychological processes but also opened new possibilities in the field of psychometrics, showing how artificial intelligence can support the process of creating psychological scales that are more intelligible, valid, and adapted to the specific needs of the studied populations.

Such reconstruction enabled not only the description of a given phenomenon in a way that would be impossible to predict purely theoretically—without precise observation—but also allowed for the identification of groups of individuals in whom the given phenomenon manifested in the most atypical way. Thanks to the use of decision trees, it became possible to classify groups of individuals according to specific attributes and process trajectories, which made it possible to identify those groups in which a particular variant of the phenomenon occurred most frequently, as well as those in which it was least typical.

Moreover, this classification made it possible not only to identify these atypical groups but also to construct specific subscales within the psychological scale tailored to them. This is a pioneering approach that has not previously been applied in psychometrics. Traditional psychological scales often average results and ignore specific cases, which leads to overlooking individuals whose experiences differ

slightly from the majority. Thanks to artificial intelligence—and decision trees in particular—it became possible to develop diagnostic tools that also account for these atypical cases.

As a result, even for individuals who previously might have been disregarded due to their atypicality, this method enables the construction of precise questions that reflect their unique experiences. This approach not only increases the validity of the psychological scale but also broadens the capacity for diagnosing and understanding psychological phenomena in their full complexity (Rzechowska & Szymańska, 2017).

CHAPTER 40

Profiling Psychological Traits in Kernel Spaces

In classical psychometric approaches, an individual's psychological profile is treated as a set of point scores on particular scales—typically visualised as a multidimensional vector in trait space. While this approach is useful, it assumes a linear space in which variables are independent and subject to simple distance metrics. In situations where the profile includes many interdependent traits, and the relationships between them are nonlinear or structurally complex, classical modeling methods prove insufficient.

In response to these limitations, Szymańska (2025b) proposed an original concept of psychometric modeling within reproducing kernel Hilbert spaces (RKHS) (Szymańska, 2025b). This approach involves transforming psychological data—using kernel functions such as the radial basis function (RBF)—into a higher-dimensional space where complex relationships between traits can be formally and geometrically captured. In this space, the psychological profile is no longer a simple vector of numbers but takes the form of a structural "shape" – a geometric configuration that can be analysed in terms of length (trait intensity), direction (pattern of dominant properties), and similarity to theoretical templates.

The model profile, in Szymańska's view, is therefore not merely a collection of scores but a function mapped into Hilbert space, which allows for the matching of empirical data to ideal theoretical representations. This approach proves particularly useful in situations where an individual's profile includes a large number of traits (e.g., 12 or 15), and the relationships among them are nonlinear and dispersed. In RKHS, such profiles become more distinguishable, classifiable, and comparable to previously defined model curves.

The core of this method lies in shifting the focus from individual measurement points to the overall geometric structure of the profile—treated as a mathematical object in kernel space. The length of the vector may be interpreted as the general psychological strength of the profile (e.g., the intensity of a cluster of traits), while its direction reflects the typological configuration. This enables not only more precise alignment of theory and data but also the detection of deviations from typical patterns, the exploration of borderline cases, and the construction of diagnostic models based on the geometric structure of the data.

The conceptualisation of kernel spaces in psychology opens new possibilities within psychometrics that were previously beyond the reach of classical statistical analyses. This includes, among others, fitting theoretical models to empirical data, detecting high-dimensional structures in personality data, and modeling diagnostic transition spaces between personality types and disorders.

CHAPTER 41

New Frontiers in Psychometrics: Artificial Intelligence, Psychocybernetics, and the Redefinition of Psychometric Paradigms

Psychology, as a science, has achieved significant success over just a century—both in conceptualising psychological traits and in developing methods for their measurement. A particular challenge has been the creation of tools enabling the quantitative determination of latent constructs—that is, traits that cannot be directly observed. Unlike variables that manifest physically, such as height or body mass, psychological traits lack a clear empirical counterpart and can only be inferred on the basis of behavioural indicators.

Every measurement in psychometrics is based on the relationship between observable variables (indicators) and latent variables (illata), which are the source of those observations. The illatum, as a theoretical construct, is not measured directly but reveals itself in the form of variables that manifest its presence. Indicators are therefore an indirect representation of latent traits, and their appropriate selection and statistical modeling form the basis for accurate psychological measurement.

A perfect example of this relationship is the measurement of intelligence. Intelligence is not directly observable or measurable as a physical entity, but its manifestations—such as reaction time, problem-solving ability, or information processing capacity—can be treated as indicators of the latent intellectual trait. In this context, the proper operationalisation of illata requires the use of psychometric models that allow for the translation of raw observations into a quantitative result corresponding to the level of the trait under study.

Psychology has managed to overcome these challenges by employing methods such as structural equation modeling (SEM), factor analysis, and item response theory (IRT), which enable precise estimation of latent traits based on observable indicators. In doing so, psychometrics has developed a coherent methodological apparatus that makes it possible to quantify psychological constructs and, thus, to study the human mind in an empirically grounded manner.

Psychometrics is therefore one of psychology's greatest achievements—not only because it made it possible to measure latent traits, but also because it enabled their comparison across individuals. This achievement was made possible through the introduction of the concept of norm, which in psychometrics takes the form of a value typical for the population—that is, the expected value.

The expected value in the statistical sense is the arithmetic mean of results within a given population, serving as a point of reference for the classification of individuals in terms of a given psychological trait. The introduction of this concept made it possible not only to measure psychological traits but also to interpret them in a normative context—allowing for the determination of whether an individual's score falls within the range of typical values or deviates from them significantly.

Thanks to this, psychology gained the ability not only to measure latent traits but also to compare individuals in terms of these traits. Psychometric measurement does not exist in a vacuum—for a result to be meaningful, it must be situated in relation to the distribution of scores within the population. It is the statistical norm, defined as the population mean, that allows for the classification of results and the drawing of conclusions about the individual in the context of the group. Although the statistical norm enables comparisons, it is, as Kurt Lewin emphasized, theory that defines what is to be measured in the first place.

As a result, psychometrics gained the ability to develop scales, standardised tests, and classification methods that made it possible to precisely determine whether an individual's score is typical or deviates from the norm—and if so, to what extent. This constitutes a key foundation of psychological measurement, which enabled the development of scientific diagnostics and the creation of models based on normative population values.

However, psychological models still lack a definition of what constitutes a norm on the individual level—not just the population level. In other words, psychology has not yet established a clear concept of norm in reference to the behaviour of the human being as an individual. Is a norm merely that which is statistically typical, or is there a need to define norm in a more fundamental way?

Contemporary psychology, relying on a statistical approach, adopts the norm as the value typical for a population, meaning that behaviour falling within a range defined by the mean and standard deviation is considered "normative". Yet in light of current trends—where more and more individuals in the population meet the criteria for depression—the question arises: if the majority of the population qualifies for a diagnosis of depression, does depression then become the norm? If the norm is merely a statistical reflection of dominant characteristics in the population, this

could lead to a paradox in which disorder becomes the norm, and what was once regarded as psychological health is treated as deviation.

This suggests that a new approach to the notion of norm is needed—one that does not refer solely to population statistics but to intrinsic value in itself. But how can psychological norm be defined as an autonomous value, independent of the population distribution? This question requires philosophical reflection: is the norm what is typical, or rather what promotes effective and balanced individual functioning?

It seems worthwhile to consider expanding the dominant normative–statistical paradigm in psychology and to explore the possibility of redefining the concept of norm in a more functional way. In this context, cybernetics may offer a new perspective in which the norm is not merely a statistical average but is related to the effectiveness and adaptability of the organism (Mazur, 1966, 1976; Szymańska, 2024b). Norm as functionality could offer a new perspective for psychometrics—enabling the assessment not only of how frequently a given behaviour occurs, but also of how well it supports effective and balanced functioning of the individual within their environment (Szymańska, 2023a).

Szymańska (2024b) raises this issue in her work, suggesting that the application of a cybernetic concept of system functionality—one that allows the system to maintain homeostasis and operate in a balanced way—could complement the existing understanding of psychological norm. The aim is not to replace the classical statistical norm, but rather to extend it with an additional perspective that could be employed in situations requiring a different approach to the analysis of individual behaviour.

Although cybernetics has long been present in psychology—particularly in the areas of regulation theory and adaptation (Tomaszewski, Łukaszewski, Kozielecki)—its assumptions appear to offer potential for redefining the very notion of the psychometric norm—as a foundation for measurement, rather than merely a descriptor of behaviour.

Statistically defined norm, as a value typical for the population, is and will remain a key reference point in psychometrics. It is precisely this norm that makes possible the comparison of individual results, the standardisation of tests, and the diagnosis of deviations from the average values in a given group. This approach is fundamental to psychology, and its role remains unquestionable. However, there are situations in which it would be both possible and beneficial to apply a different definition of norm—one that does not refer to typicality, but rather to the functionality of the individual within their environment.

In the proposed framework, existing concepts of norm could be extended by introducing a second interpretation—based on the principles of homeostasis and balanced system functioning, drawn from the cybernetic tradition (Szymańska, 2024b). In this approach, normativity would not result from comparison with the population but from an evaluation of the individual's capacity to maintain psychological equilibrium and act effectively.

The concept of homeostasis was first applied by W. Cannon in reference to physiological systems, and later developed by P. Anokhin in the context of the organism's functional systems (Egiazaryan & Sudakov, 2007). These classical approaches demonstrate that the logic of regulation and the system's ability to maintain equilibrium are universal in nature and lend themselves to formal modeling. Therefore, it is both natural and consistent to extend this logic to the psychometric level—where statistical and descriptive models have traditionally prevailed. In this sense, the author's proposal represents a continuation of this tradition, but transferred into the domain of psychometrics—as a foundation for measurement, not merely a biological description of system functioning.

Such a dual system of normativity could find application in various psychological contexts. When the goal is to compare individuals with one another, the norm of typicality should be applied, as it enables the classification of results in relation to the general population. However, in situations where we are interested in the individual's capacity for balanced and effective functioning, the functional norm should be employed—assessing not only whether behaviours are typical, but above all, whether they enable effective adaptation to the environment and the maintenance of psychological stability.

Introducing such a distinction could revolutionise the way normativity is understood in psychology. It would open new possibilities in diagnostics, allowing the interpretation of psychometric results to be adapted to the purpose of the assessment—in some cases based on the statistical norm, and in others on the evaluation of the individual's ability to regulate and stabilise their psychological system.

Of course, the issue of functionality in psychology remains an open question and requires further development. Cybernetic models, such as Mazur's autonomous system, provide precise tools for describing mechanisms of psychological regulation that could serve as a foundation for a new conception of norm—as the optimal organisation of the psychological system.

Szymańska (2023a) has already demonstrated how the functionality of the psychological system is disrupted in the case of personality disorders such as depressive personality. In this model, the homeostasis of the system is impaired, leading to a loss of capacity for effective self-regulation and adaptation. In depressive personality, the psychological system loses its ability to maintain stability, which manifests, among other things, in social withdrawal, decreased energy, and reduced responsiveness to environmental stimuli.

Similar mechanisms can be traced in other personality structures. An antisocial individual may exhibit a breakdown of functionality through severing connections with the environment, whereas a histrionic individual may demonstrate dysfunction through an inability to maintain internal equilibrium and a compulsive drive for external emotional validation. Cybernetic models can offer a precise description of how such functional disorders occur and which regulatory mechanisms fail across different personality structures.

This approach opens a new perspective in psychometrics and psychological diagnostics. Instead of focusing solely on comparing the individual to the population,

it enables the analysis of how the organisation of a person's psychological system affects their ability to maintain equilibrium and function effectively. Although the proposed approach draws upon classical cybernetic logic, its novelty lies in applying this logic as the basis for constructing a psychometric norm—something that has not yet been developed in the diagnostic literature.

It is therefore worth considering a deeper investigation into the cybernetic understanding of functionality in relation to psychological norm—especially in the context of diagnostics and the construction of psychometric tests. While grounded in the classical tradition of system regulation, this approach may offer an inspiring direction for further development in thinking about psychological measurement. In the future, models based on systems theory could become a key tool in assessing mental health, enabling not only the classification of disorders but also the prediction of their development and the selection of intervention strategies. This is a direction that—if further developed—could significantly enrich the way psychological norm and its measurement are conceptualised.

SUMMARY

The aim of this book was to familiarise the reader with the issues of mathematical modeling in psychology using artificial intelligence, demonstrating how modern algorithms can support data analysis, model construction, and the prediction of psychological phenomena. Both classical statistical methods and advanced AI techniques were discussed, highlighting their application in psychometrics, language analysis, modeling of cognitive and behavioural processes, and psychological diagnostics. The interdisciplinary nature of these topics was emphasised, showing that mathematics, statistics, and algorithms can be effectively applied in psychological research.

One of the main messages of the book is the presentation of artificial intelligence as a set of computational procedures rooted in precise mathematical and statistical principles. AI is not an autonomous entity, but a tool that requires deliberate design, implementation, and interpretation of results. Its effective application in psychology demands solid methodological grounding—any analysis that lacks theoretical foundations carries the risk of errors and misinterpretations. AI tools can enhance psychological research, provided that users understand not only their potential but also the limitations arising from the assumptions on which they are based.

The book demonstrated how algorithms can support psychologists in analysing complex data structures, enabling deeper understanding of psychological processes and their dynamic interrelations. Particular attention was devoted to models analysing cognitive, emotional, and social processes. It was underscored that interpreting results obtained through AI cannot be limited to reading numerical indicators—it also requires a sound understanding of the psychological foundations of the phenomena being studied. Artificial intelligence in psychology becomes a breakthrough tool, but only when its use remains critical, contextual, and consciously adapted to the nature of the data being analysed.

A key emphasis of the book was the assertion that knowledge of artificial intelligence must not be detached from its mathematical foundations. Effective use of modern analytical tools requires expertise in statistics, probability theory, and modeling. Every algorithm is based on specific assumptions that directly affect the outcomes—ignoring these dependencies may lead to erroneous interpretations of data. Therefore, the need to combine technical competence with psychological sensitivity and interpretative skill was strongly emphasised.

The book does not shy away from reflecting on the limitations and challenges associated with the application of AI in psychology. One of the most significant issues remains the quality of input data—psychological data are often incomplete, subjective, and burdened with measurement error, which may lead to distorted results. Moreover, the validation of AI-based models requires particular caution—not all algorithms perform equally well in the analysis of human behaviour, and overinterpretation of their results can lead to erroneous diagnoses or research conclusions.

The appropriate interpretation of results generated by mathematical models and AI algorithms requires careful consideration of the context in which the data were collected and processed. Psychology cannot afford to reduce the complexity of human behaviour to purely numerical patterns. For this reason, AI should not replace the psychologist but rather support their work—providing data, assisting in analysis, but not assuming the interpretative role. It is crucial to maintain a balance between technological capabilities and an understanding of psychological complexity.

Within this context, the book presents selected algorithms and mathematical modeling techniques that, in the author's view, offer the greatest application potential in psychology and psychometrics. The goal was not to exhaustively discuss all possible approaches, but to outline a foundation for the further development of interdisciplinary methods of data analysis. Many of the directions described require further research and practical implementation, which remains an open field for exploration.

Particular attention was devoted to the topic of psychological profiling and its potential applications in psychometrics and forensic science. It was emphasised that traditional approaches based on intuition and expert experience are difficult to validate objectively. Artificial intelligence may introduce a new standard here, offering formalised, measurable models of psychological profiles. Such an approach increases replicability, enables the verification of validity, and creates new opportunities in forensic assessment, clinical psychology, and the study of human behaviour.

One of the key elements of the profiling process is the matching of model curves—a method that allows for the evaluation of the fit between empirical data and model-based patterns. The book demonstrates how a mathematical approach to this issue helps eliminate subjective components of interpretation and increases analytical precision. A particularly promising direction for further research is the automatic calibration and adaptation of curves, which may find application in the analysis of intelligence, emotions, and cognitive processes.

Another prospective area of development is three-dimensional modeling of psychological traits. Classical psychometric methods operate in two-dimensional systems, which limits the ability to capture complex relationships between personality traits, emotional processes, and cognitive mechanisms. Introducing higher dimensions enables more precise mapping of the structure of psychological constructs, opening new possibilities in psychometrics as well as in modeling the dynamics of emotion and behaviour.

The book also highlights the potential of psychocybernetics as a bridge between mathematical modeling and classical psychological theories. Viewing the mind as a control and information-processing system offers a new perspective on regulatory and adaptive processes. Within this context, the problem of norm in psychology—long discussed yet difficult to capture empirically—takes on particular significance. Cybernetic models appear to make its operationalisation possible, which could translate into more objective diagnostic and therapeutic tools.

The book further emphasises the diversity of artificial intelligence algorithms and their complementary applications in psychology. AI is not limited to language models—although models such as GPT, BERT, or DeepSeek have gained popularity for their ability to generate text and simulate human communication. In reality, they represent only a segment of the broader analytical landscape. While their application is important in psycholinguistics and narrative analysis, they do not dominate over other algorithms used in psychometrics or data modeling. The book situates language models within a wider context, demonstrating that their true strength emerges only when integrated with other analytical approaches.

From this perspective, AI emerges as a multidimensional set of tools enabling more precise modeling of complex psychological processes. Understanding this diversity allows for a more accurate evaluation of how artificial intelligence can support psychology—not by simplifying it, but by enriching it with new methods of analysis and prediction.

It is my hope that this book will help dispel the perception of AI as an enigmatic technology accessible only to computer scientists, and instead present it as an accessible and practical tool in the hands of psychologists. Contemporary science cannot thrive without an interdisciplinary approach—mathematics and statistics are the foundation upon which the effective application of AI in psychology is built. The issues discussed here are intended not only to introduce methods, but also to change the way we think—so that AI is no longer seen as an inaccessible "black box", but rather as a set of tools that, when used consciously and responsibly, can revolutionise psychology, opening new avenues for research and application.

At the close of this book, it is worth asking: what is it all for? Why all this effort to delve into mathematical models, artificial intelligence, and psychometrics? Is it only about more accurate diagnoses, better tools, more refined analyses?

Not only.

Throughout this book, we have spoken about profiling—about creating multidimensional, precise descriptions of the individual. We have learned how mathematics and psychology can together shape an image of the human being within data structures and trait spaces. But perhaps this profile is not merely a tool of understanding. Perhaps it is something more—a foundation for a new kind of relationship: between human and technology.

A new generation of systems is emerging today—AI agents that are no longer merely tools for analysis, but digital entities designed for close coexistence with human beings. Their role is to support, accompany, and at times even protect. But to do so effectively, they must know us deeply—not as data points, but as unique coordinates in a psychological space.

Such an agent is more than just an application. It is a personal system—customised, secure, encoded within a decentralised structure (Web3, blockchain)—private, inviolable, exclusively ours. It does not operate based on generalised patterns but, drawing on psychological knowledge, recognises our needs, moods, and habits. It knows we like soft socks and soft-boiled eggs, but also that after a sleepless night we need silence. And not because someone programmed it that way—but because it has learned who we are.

This vision becomes particularly vital in the context of individuals with neurocognitive disorders. For someone who forgets, who loses orientation in time and space, such an agent could serve as a cognitive prosthesis—not of a limb, but of the mind. It could remember, remind, recognise emotions, manage the household, and communicate with a physician. It would offer real support—not as a substitute for a human being, but as an extension rooted in true understanding. Not feigned empathy, but action based on genuine knowledge.

And this is where psychology reaches beyond itself. It is no longer merely the science of the human being—it becomes knowledge capable of guiding technology. AI was partly born from psychological inspirations. Now the moment has come when psychology can give something back to AI—not a set of tests, but deep insight. The ability to discern what makes us unique. To build agents that are not anonymous, but embedded in intimate, personal recognition.

Perhaps this is the most important message of this book: what we are learning today—these models, these data, these theories—are not merely intellectual tools. They are bricks. The first, and most essential. Laying the foundation for a future in which technology not only supports us but truly knows us—and understands that we are unrepeatable.

And perhaps that is why psychological profiling, which has been at the centre of this book, matters so deeply today. Because only it can teach artificial intelligence to see the human being in a way that truly makes sense.

To AI systems, a human is not a "person" in the way we understand it—not someone with a scent, a gaze, a voice that trembles with emotion. AI perceives differently. What it sees is a point in space—a constellation of traits, relationships,

values. But not dead data. A structure. A dynamic network of tensions, variables, and links that forms something singular.

This is precisely why it needs psychology. It needs it to understand which traits matter. Which dimensions are key. Which interrelations form a human being. Psychology gives it the language to capture sensitivity. Precision. Uniqueness.

Without psychology, artificial intelligence may be fast and effective—but never personal. Never truly attuned. Without it, AI sees datasets, not people. Zosia, seen by AI, is merely a string of numbers—until it is given a map. And it is psychology that provides this map.

Thanks to psychology, an AI agent can perceive a person not as a statistical profile, but as a unique point in a vast space of traits—irreplaceable, inimitable. It can come to understand that it is not interacting with a "user", but with *this specific individual*—with her habits, emotions, needs, and history.

This is no longer just the technology of the future. It is happening now.

BIBLIOGRAPHY

Abramek, E., & Rizun, M. (2015). Wykorzystanie analizy sieci społecznych do badania kapitału intelektualnego na przykładzie platformy e-learningowej. *Innowacje w Zarządzaniu i Inżynierii Produkcji (Innovations in Production Management and Engineering), Zakopane, Poland, Volume: II*, 11–25. http://46.242.185.119/off_ptzp.org.pl/files/konferencje/kzz/artyk_pdf_2015/T2/t2_0011.pdf

Alamsyah, A., Rahardjo, B., & Kuspriyanto. (2013). Social Network Analysis Taxonomy Based on Graph Representation. *The 5th Indonesian International Conference on Innovation, Entrepreneurship, and Small Business (IICIES)*, *June*, 341–349. https://doi.org/10.13140/2.1.3221.2160

Alawadh, M. M., & Barnawi, A. M. (2022). A Survey on Methods and Applications of Intelligent Market Basket Analysis Based on Association Rule. *Journal on Big Data Tech Science*. https://doi.org/10.32604/jbd.2022.021744

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., … Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

American Psychological Association. (2020). *APA Dictionary of Psychology*. American Psychological Association.

Anastasi, A., & Urbina, S. (1999). *Testy psychologiczne*. Pracownia Testów Psychologicznych Polskiego Towarzystwa Psychologicznego.

Apostolato, I. A. (2013). An overview of Software Applications for Social Network Analysis. *International Review of Social Research*, *3*(3), 71–77. https://doi.org/10.1515/irsr-2013-0023

Aranowska, E. (1989). Wskaźniki bliskości dowolnej krzywej do krzywej modelowej. *Psychologia Matematyczna*, *3*, 99–114.

Aranowska, E. (1996). *Metodologiczne problemy zastosowań modeli statystycznych w psychologii. Teoria i praktyka [Methodological problems of the use of statistical models in psychology. Theory and practice]*. Studio 1.

Aranowska, E. (2005). *Pomiar ilościowy w psychologii [Quantitative measurements in psychology]*. Wydawnictwo Naukowe SCHOLAR.

Aranowska, E., & Rytel, J. (2010). Wielowymiarowa analiza wariancji – MANOVA. *Psychologia Społeczna*, *3*(14), 117–141.

Aranowska, E., & Rytel, J. (2013). Kontrowersje wokół rzetelności jako pojęcia psychometrycznego. *Przegląd Psychologiczny: organ Polskiego Towarzystwa Psychologicznego*, *1*(56), 29–43. Aranowska, E., & Szymańska, A. (2017). Trafność zmiennej latentnej wyjaśnianej przez model SEM. *Ogólnopolska Konferencja Naukowa "Katowickie Spotkania Psychometryczne"*.

Arif, T. (2015). The Mathematics of Social Network Analysis: Metrics for Academic Social Networks. *International Journal of Computer Applications Technology and Research*, *4*(12), 889–893. https://doi.org/10.7753/ijcatr0412.1003

Babbie, E. (2007). *Badania społeczne w praktyce*. Wydawnictwo Naukowe PWN.

Barbaro, B. De. (1999). *Wprowadzenie do systemowego rozumienia rodziny*. Wydawnictwo Uniwersytetu Jagiellońskiego.

Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in Neurobiology*, *192* (October 2019). https://doi.org/10.1016/j.pneurobio.2020.101821

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data*. Chapman & Hall/CRC Press.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 – Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Berg, R. A. (2008). *Statistical learning from a regression perspective*. Springer-Verlag.

Biela, A. (1995). *Skalowanie wielowymiarowe w analizach ekonomicznych i behawioralnych*. Norbertinum.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(3), 993–1022. https://doi.org/10.1016/B978-0-12-411519-4.00006-9

Blunsom, P. (2004). *Hidden markov models*. https://cs.wmich.edu/~alfuqaha/Fall11/cs6570/lectures/hmm-tutorial.pdf

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C. I., Gómez-Gardenes, J., Romance, M., Wang, Z., & Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, *544*(1), 1–122.

Bokus, B., Bartczak, M., Szymańska, A., Chronowska, R., & Ważyńska, A. (2017). The Dialogical Self's Round Table: Who Sits at It and Where? *Psychology of Language and Communication*, *21*(1), 84–108.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020–Decem*.

Brzezińska, A. (2002). *Społeczna psychologia rozwoju [Social psychology of development]*. Wydawnictwo Naukowe SCHOLAR.

Brzeziński, J. (1996). *Metodologia badań psychologicznych*. Wydawnictwo Naukowe PWN.

Brzeziński, J., & Stachowski, R. (1984). *Zastosowanie analizy wariancji w eksperymentalnych badaniach psychologicznych*. Państwowe Towarzystwo Psychologiczne.

Burke, N., Brezack, N., & Woodward, A. (2022). Children's social networks in developmental psychology: A network approach to capture and describe early social environments. *Frontiers in Psychology*, *13*(October). https://doi.org/10.3389/fpsyg.2022.1009422

Burt, R. S. (2022). Structural holes and good ideas. *Handbook of Sociological Science: Contributions to Rigorous Sociology*, *110*(2), 372–422. https://doi.org/10.4337/9781789909432.00030

Butts, C. T., Leslie-Cook, A., Krivitsky, P. N., Bender-deMoll, S., Almquist, Z., Hunter, D. R., Wang, L., Li, K., Goodreau, S. M., Horner, J., & Morris, M. (2024). *NetworkDynamic: Dynamic Extensions for Network Objects* (R package version 0.11.5). https://cran.r-project.org/web/packages/networkDynamic/

Bzdok, D., & Meyer-Lindenberg, A. (2017). *Machine learning for precision psychiatry*. 1–16. http://arxiv.org/abs/1705.10553

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649–685. https://doi.org/10.1017/S1351324916000383

Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). *Spanish Pre-trained BERT Model and Evaluation Data*, 1–9. http://arxiv.org/abs/2308.02976

Chakraborty, A., Dutta, T., Mondal, S., & Nath, A. (2018). Application of Graph Theory in Social Media. *International Journal of Computer Sciences and Engineering*, *6*(10), 722–729. https://doi.org/10.26438/ijcse/v6i10.722729

Chang, K. H. (2023). Natural Language Processing: Recent Development and Applications. *Applied Sciences (Switzerland)*, *13*(20), 10–11. https://doi.org/10.3390/app132011395

Chen, J., & Lv, S. (2022). Long Text Truncation Algorithm Based on Label Embedding in Text Classification. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199874

Chen, Y. J., Jhang, K. M., Wang, W. F., Lin, G. C., Yen, S. W., & Wu, H. H. (2022). Applying Apriori algorithm to explore long-term care services usage status—Variables based on the combination of patients with dementia and their caregivers. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1022860

Chimieski, B., & Fagundes, R. (2013). Association and Classification Data Mining Algorithms Comparison over Medical Datasets. *Journal of Health Informatics*, *5*(2), 44–51. Retrieved from https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/226

Cholewa, W. (1995). *Elementy systemów doradczych wspomagających badania diagnostyczne maszyn. Prace Naukowe Instytutu Technologii Maszyn i Automatyzacji Politechniki Wrocławskiej*, 57(23).

Chopra, A., Prashar, A., & Sain, C. (2013). Natural language processing. *International Journal of Technology Enhancements and Emerging Engineering Research*, *1*(4), 131–134.

Choynowski, M. (1971). Podstawy i zastosowania teorii testów psychologicznych. In J. Kozielecki (Ed.), *Problemy psychologii matematycznej* (pp. 65–118). Państwowe Wydawnictwo Naukowe.

Cierpka, A. (2004). Tożsamość jednostki a narracje rodzinne – propozycje badawcze. In T. Maruszewski (Ed.), Adaptacja do zmian. *Kolokwia Psychologiczne 12* (pp. 91–105). Wydawnictwo Instytutu Psychologii PAN.

Ciok, A. (2004a). Asymmetry and the inverse concentration set. In T. Kowalczyk, E. Pleszczyńska, & F. Ruland (Eds.), *Grade Models and Methods for Data Analysis* (pp. 139–165). Springer.

Ciok, A. (2004b). Discretization and regularity. In T. Kowalczyk, E. Pleszczyńska, & F. Ruland (Eds.), *Grade Models and Methods for Data Analysis* (pp. 167–184). Springer.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*, 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2025). *The rising costs of training frontier AI models. 2016*, 1–20. http://arxiv.org/abs/2405.21015

Dash, C. S. K., Behera, A. K., Dehuri, S., & Cho, S. B. (2016). Radial basis function neural networks: A topical state-of-the-art survey. *Open Computer Science*, *6*(1), 33–63. https://doi.org/10.1515/comp-2016-0005

Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, *21*, 17–23. https://doi.org/10.3354/esep00195

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dornyei, Z., & Otto, I. (1998). *Motivation in action: A process model of L2 motivation*. *4*, 43–69. http://eprints.nottingham.ac.uk/39/

Dramiński, M. (2007). *Algorytm indukcji reguł decyzyjnych w problemach klasyfikacji i wyboru cech w zadaniach wyskowymiarowych*. Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Du, K. L. (2010). Clustering: A neural network approach. *Neural Networks*, *23*(1), 89–107.

Duch, W., Korbicz, J., Rutkowski, L., & Tadeusiewicz, R. (Eds.). (2000). *Sieci neuronowe* (*Biocybernetyka i inżynieria biomedyczna 2000*, Vol. 6). Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Egiazaryan, G. G., & Sudakov, K. V. (2007). Theory of functional systems in the Scientific School of P. K. Anokhin. *Journal of the History of the Neurosciences*, *16*(1–2), 194–205. https://doi.org/10.1080/09647040600602805

Elder, J., Hill, T., Miner, G., Nisbet, B., Delen, D., & Fast, A. (2012). *Practical Text Mining and Statistical Analysis for Nono-structured Text Data Application*. Elsevier.

Esposito, M., Masala, G. L., Minutolo, A., & Pota, M. (2021). Special issue on "natural language processing: Emerging neural approaches and applications". *Applied Sciences (Switzerland)*, *11*(15). https://doi.org/10.3390/app11156717

Evgeniou, T., & Pontil, M. (2001). Workshop on Support Vector Machines: Theory and Applications. *Support Vector Machines: Theory and Applications*. https://doi.org/10.1007/3-540-44673-7_12

Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science* (Issue January 2004). Empirical Press.

Friedman, N. (2013). The Bayesian Structural EM Algorithm. *Arxiv*, 129–138. http://arxiv.org/abs/1301.7373

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text Mining Methods and Techniques. *International Journal of Computer Applications*, *85*(17), 42–45. https://doi.org/10.5120/14937-3507

Gajda, J. (1992). Modele strukturalne w naukach społecznych. In E. Aranowska (Ed.), *Wybrane problemy metodologii badań* (pp. 100–132). Wydawnictwa Uniwersytetu Warszawskiego.

Gales, M., & Young, S. (2007). The application of hidden Markov Models in speech recognition. *Foundations and Trends in Signal Processing*, *1*(3), 195–304. https://doi.org/10.1561/2000000004

Ganati, G. A., Repalle, V. N. S. R., & Ashebo, M. A. (2023). Social network analysis by Turiyam graphs. *BMC Research Notes*, *16*(1), 1–10. https://doi.org/10.1186/s13104-023-06435-7

Garson, G. D. (2013). Introductory guide to HLM with HLM 7 software. In *Hierarchical linear modeling: Guide and applications* (pp. 55–96). Sage.

Gaudart, J., Giusiano, B., & Huiart, L. (2003). Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. *Computational Statistics & Data Analysis*, *44*(4), 547–570.

Gaul, M., & Machowski, A. (1987). Elementy analizy ścieżek. In J. Brzeziński (Ed.), *Wielozmiennowe modele statystyczne w badaniach psychologicznych*. Państwowe Wydawnictwo Naukowe.

Geyer, P. (2012). Extraversion – Introversion: what C. G. Jung meant and how contemporaries responded. *AusAPT National Conference*, *October 2012*. https://www.researchgate.net/publication/264782791_Extraversion_-_Introversion_what_CG_Jung_meant_and_how_contemporaries_responded?enrichId=rgreq-2447f6b7372a1d65927867c3994df79d-XXX&enrichSource=Y292ZXJQYWdlOzI2NDc4Mjc5MTtBUzoxMzA2NjI4NzkyMDc0MjRAMTQwODE

Glenn, E. (2005). Incorporating parental goals in parenting programs through collaborative relationships with parents. *Journal of Extension*, *43*(1).

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026

Goh, T. T., Jamaludin, N. A. A., Mohamed, H., Ismail, M. N., & Chua, H. (2023). Semantic Similarity Analysis for Examination Questions Classification Using WordNet. *Applied Sciences*, *13*(14). https://doi.org/10.3390/app13148323

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Grzesiuk, L. (2005). Dynamika relacji pacjent–terapeuta. In L. Grzesiuk (Ed.), *Psychoterapia. Teoria* (pp. 355–384). ENETEIA.

Grzesiuk, L., Szymańska, A., & Dobrenko, K. (2017). Znaczenie pracy nad przeniesieniem dla związku między zahamowaniem pacjenta przed psychoterapią a dobrą relacją z psychoterapeutą i skutecznością psychoterapii. In D. Danielewicz & J. Rola (Eds.), *Stare dylematy i nowe wyzwania w psychoterapii* (pp. 247–261). Wydawnictwo Akademii Pedagogiki Specjalnej.

Gupta, S., Pandey, S., & Shukla K. K. (2015). Comparison Analysis of Link Prediction Algorithms in Social Network. *International Journal of Computer Applications*, *111*(16), 27–29. https://doi.org/10.5120/19624-1502

Gurycka, A. (1990). *Błąd w wychowaniu [Mistake in upbringing]*. Wydawnictwa Szkolne i Pedagogiczne.

Gurycka, A. (2008). Błędy w wychowaniu [Mistakes in upbringing]. In E. Kubiak-Szymborska & D. Zając (Eds.), *O wychowaniu i jego antynomiach [On upbringing and its antinomies]*. Wydawnictwo WERS.

Hair, J. J., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle River, NJ: Pearson.

Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. (2020). Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1), 1–34. https://doi.org/10.3390/bdcc4010001

He, L., Zhang, Q., Duan, J., & Wang, H. (2023). An Open-Domain Event Extraction Method Incorporating Semantic and Dependent Syntactic Information. *Applied Sciences (Switzerland)*, *13*(13). https://doi.org/10.3390/app13137942

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28.

Heck, R. H., & Thomas, S. L. (2009). *An introduction to multilevel modeling techniques*. Routledge.

Herff, C., & Schultz, T. (2016). Automatic speech recognition from neural signals: A focused review. *Frontiers in Neuroscience*, *10*(SEP), 1–7. https://doi.org/10.3389/fnins.2016.00429

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign. http://bit.ly/AIED- BOOK%0A

Hornowska, E. (2003). *Testy psychologiczne: teoria i praktyka*. Wydawnictwo Naukowe SCHOLAR.

Hoque, M. D. J., Islam, M. S., & Mohtasim, S. A. (2024). Optimizing Decision-Making Through Customer-Centric Market Basket Analysis. *Journal of Operational and Strategic Analytics*, *2*(2), 72–83. https://doi.org/10.56578/josa020201

Hüsken, M., & Stagge, P. (2003). Recurrent neural networks for time series classification. *Neurocomputing*, *50*, 223–235. https://doi.org/10.1016/S0925-2312(01)00706-8

Jarochowska, E. (2005a). Analiza danych, dotyczących postrzegania własnego zdrowia w programie GradeStat. In J. B. Książyk, O. Matyja, E. Pleszczyńska, & M. Wiech (Eds.), *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat* (pp. 9–42). Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Jarochowska, E. (2005b). Na czym opiera sie gradacyjna analiza danych? Intuicyjne wprowadzenie zamiast teorii. In *Analiza danych medycznych i demograficznych przy użyciu programu GradeStat* (pp. 215–236). Instytut Podstaw Informatyki Polskiej Akademii Nauk.

Jaworowska, K., Szymańska, A., Bartczak, M., & Bokus, B. (2016). Metaphorical conceptualization of notion: the role of mood. In H. Kyuchukor (Ed.), *New Trends in the Psychology of Language* (pp. 83–106). LINCOM GmbH.

Jeon, K. C., & Goodson, P. (2015). *US adolescents' friendship networks and health risk behaviors: A systematic review of studies using social network analysis and Add Health data.* PeerJ, *3*, e1052. https://doi.org/10.7717/peerj.1052

Jha, J., & Ragha, L. (2013). Educational Data Mining using Improved Apriori Algorithm. *International Journal of Information and Computation Technology*, *3*(5), 411–418. Retrieved from https://www.ripublication.com/irph/ijict_spl/08_ijictv3n5spl.pdf

Jiang, X., Chen, Y., Ao, N., Xiao, Y., & Du, F. (2022). A Depression-Risk Mental Pattern Identified by Hidden Markov Model in Undergraduates. *International Journal of Environmental Research and Public Health*, *19*(21). https://doi.org/10.3390/ijerph192114411

Jonkisz, A. (1998). *Ciągłość teoretycznych wytworów nauki. Ujęcie strukturalne [Continuity of theoretical scientific productions. A structural approach]*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Jöreskog, K. G. (1999). *How Large Can a Standardized Coefficient be?* http://www.statmodel.com/download/Joreskog.pdf

Karthiyayini, R., & Balasubramanian, R. (2016). Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis. *International Journal of Advanced Research in Computer Science and Software Engineering*, *6*(10), 2277. www.ijarcsse.com

Khan, B. S., & Niazi, M. A. (2017). Network Community Detection: A Review and Visual Survey. *ArXiv Preprint ArXiv:1708.00977*. http://arxiv.org/abs/1708.00977

Kinnear, P. R., & Gray C. D. (2008). *SPSS 15 made simple*. Psychology Press Taylor & Francis Group.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, *2*(3), 203–271. https://doi.org/10.1093/comnet/cnu016

Koczkodaj, W. W., Kakiashvili, T., Szymańska, A., Montero-Marin, J., Araya, R., Garcia-Campayo, J., Rutkowski, K., & Strzałka, D. (2017). How to reduce the number of rating scale items without predictability loss? *Scientometrics*, 1–13. https://doi.org/10.1007/s11192-017-2283-4

Konarski, R. (2009). *Modele równań strukturalnych [Structural equation models]*. Wydawnictwo Naukowe PWN.

Kowalczyk, T., Pleszczyńska, E., & Ruland, F. (2004). *Grade Models and Methods for Data Analysis*. Springer.

Kozak, J., & Juszczuk, P. (2016). Algorithms for constructing decision trees for predicting the effectiveness of the bank's telemarketing campaign. *Zeszyty Naukowe Uniwersytetu Szczecińskiego. Studia Informatica*, *39*(1), 49–59. https://doi.org/10.18276/si.2016.39-05

Kucharski, T., & Gomula, J. (1998). *Wprowadzenie do kwestionariusza MMPI Wiskad* (pp. 1–171). Pracownia Psychologii Klinicznej i Rozwoju Osobowości w Toruniu.

Kudriavtsev, M., Bezbradica, D. M., McCarren, D. A., Roantree, M., & Ngo, V. (2023). *Exploring the Trie of Rules: a fast data structure for the representation of association rules*. https://doi.org/10.1007/s10844-024-00899-0

Kulkarni, A. R., & Mundhe, S. D. (2016). A Theoretical Review on Text Mining: Tools, Techniques, Applications and Future Challenges. *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, *4*(11), 19225–19230.

Lachowska, B. (2008). Model kołowy systemu małżeńskiego i rodzinnego w opracowaniu D. H. Olsona i współpracowników oraz narzędzia jego pomiaru. *Roczniki Teologiczne*, *55*(10), 189–207.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: a Lite Bert for Self-Supervised Learning of Language Representations. *8th International Conference on Learning Representations, ICLR 2020*, 1–17.

Latif, S., Qayyum, A., Usman, M., & Qadir, J. (2018a). Cross lingual speech emotion recognition: Urdu vs. Western Languages. *Proceedings – 2018 International Conference on Frontiers of Information Technology, FIT 2018*, 88–93. https://doi.org/10.1109/FIT.2018.00023

Latif, S., Rana, R., Younis, S., & Epps, J. (2018b). Cross Corpus Speech Emotion Classification – An Effective Transfer Learning Technique. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2018–Septe*(January), 257–261. https://doi.org/10.21437/Interspeech.2018-1625

Lee, S., & Buzby, M. (2021). *Mathematical Modeling and Simulation with MATLAB Item Type Book*. http://hdl.handle.net/11122/12246

Lekkas, D., Gyorda, J. A., Moen, E. L., & Jacobson, N. C. (2022). Using passive sensor data to probe associations of social structure with changes in personality: A synthesis of network analysis and machine learning. *PLoS ONE*, *17*(11 November), 1–22. https://doi.org/10.1371/journal.pone.0277516

Lenkiewicz, S. (2012). Gradacyjna analiza danych – idea i przykład zastosowania. *Zeszyty Naukowe Wydziału Informatycznych Technik Zarządzania Wyższej Szkoły Informatyki Stosowanej i Zarządzania „Współczesne Problemy Zarządzania"*, *1*, 63–98.

Letouche, S., & Wille, B. (2022). Connecting the Dots: Exploring Psychological Network Analysis as a Tool for Analyzing Organizational Survey Data. *Frontiers in Psychology*, *13*(May), 1–11. https://doi.org/10.3389/fpsyg.2022.838093

Lewin, K. (1952). *Field theory in social sciences*. Tavistock Publications.

Liddy, E. D. (2021). Natural Language Processing. In *In Encyclopedia of Library and Information Science*. https://surface.syr.edu/istpub/63/

Lingoes, J. C. (1977). Identyfying Directions in the Space for Interpretation. In *Geometric Representations of Relational Data* (pp. 103–112). Mathesis Press.

Liu, T., Gao, J., Ni, W., & Zeng, Q. (2023). A Multi-Granularity Word Fusion Method for Chinese NER. *Applied Sciences (Switzerland)*, *13*(5), 1–15. https://doi.org/10.3390/app13052789

Liu, Y., & Wei, J. (2023). A Systematic Review of Data Mining Studies in Parenting Research. *Journal of Educational Technology Development and Exchange*, *16*(2), 126–144. https://doi.org/https://doi.org/10.18785/jetde.1602.08

Lovins, J. B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, *11*(1), 22–31.

Luger, G. F., & Stubbledfield, W. A. (1989). *Artificial Inteligence and the Design of Expert Systems*. The Benjamin/Cummings Publishing Company, Inc.

Luo, J., Jeon, M., Lee, M., Ho, E., Pfammatter, A. F., Shetty, V., & Spring, B. (2022). Relationships between changing communication networks and changing perceptions of psychological safety in a team science setting: Analysis with actor-oriented social network models. *PLoS ONE*, *17*(8 August), 1–24. https://doi.org/10.1371/journal.pone.0273899

Łozińska-Piekarska, A., & Dąbrowski, T. (2023). Profilowanie kryminalne jako nowoczesna i skuteczna technika zwalczania przestępczości i sposób poszerzenia kompetencji pracowników zatrudnionych w organach ścigania i w wymiarze sprawiedliwości. *Edukacja Ustawiczna Dorosłych*, 123–138. https://doi.org/10.34866/3787-k277

Ma, X., & Sayama, H. (2015). *Mental disorder recovery correlated with centralities and interactions on an online social network*. PeerJ, 3, e1163. https://doi.org/10.7717/peerj.1163

Malik, H. A. M. (2022). Complex network formation and analysis of online social media systems. *CMES – Computer Modeling in Engineering and Sciences*, *130*(1). https://doi.org/10.32604/cmes.2022.018015

Manning, C., & Schütze, H. (2002). Foundations of statistical natural language processing. *ACM SIGMOD Record*, *31*(3), 37. https://doi.org/10.1145/601858.601867

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). CamemBERT: A tasty French language model. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 7203–7219. https://doi.org/10.18653/v1/2020.acl-main.645

Matyja, O. (2004). The GradeStat Program. In T. Kowalczyk, E. Pleszczyńska, & F. Ruland (Eds.), *Grade Models and Methods for Data Analysis* (pp. 455–458). Springer.

Mazur, M. (1966). *Cybernetyczna teoria układów samodzielnych*. Państwowe Wydawnictwo Naukowe.

Mazur, M. (1976). *Cybernetyka i charakter*. Państwowy Instytut Wydawniczy.

Meftah, A., Seddiq, Y., Alotaibi, Y., & Selouani, S. A. (2018). Cross-corpus Arabic and English emotion recognition. *2017 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2017*, *December 2018*, 377–381. https://doi.org/10.1109/ISSPIT.2017.8388672

Messick, S. (1995). Validity of psychological assessment. Validation of inferences of persons' responses and performances as scientific inquiry into score meaning American Psychologist, 50, 741–749. https://doi.org/10.1037/0003-066x.50.9.741

Meštrović, A., Petrović, M., & Beliga, S. (2022). Retweet Prediction Based on Heterogeneous Data Sources: The Combination of Text and Multilayer Network Features. *Applied Sciences (Switzerland)*, *12*(21). https://doi.org/10.3390/app122111216

Michalik, K. (2006a). *NEURONIX Symulator Sztucznych Sieci Neuronowych*. Artificial Intelligence Laboratory.

Michalik, K. (2006b). *PC-Shell szkieletowy system ekspertowy*. AITECH.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 – Workshop Track Proceedings*, 1–12.

Miller, R. (1966). *Proces wychowania i jego wyniki [The upbringing process and its results]*. Biblioteka Nauczyciela PZWS.

Millon, T., & Davis, R. (1996). *Disorders of Personality: DSM-IV and Beyond* (2nd ed.). John Wiley and Sons.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Mohammed, M. N., Al Dallal, A., Emad, M., Emran, A. Q., & Al Qaidoom, M. (2024). A Comparative Analysis of Artificial Hallucinations in GPT-3.5 and GPT-4: Insights into AI Progress and Challenges. In E. AlDhaen, A. Braganza, A. Hamdan, & W. Chen (Eds.), *Business Sustainability with Artificial Intelligence (AI): Challenges and Opportunities* (pp. 197–203). Springer. https://doi.org/10.1007/978-3-031-71318-7_18

Montejo-Ráez, A., & Jiménez-Zafra, S. M. (2022). Current Approaches and Applications in Natural Language Processing. *Applied Sciences (Switzerland)*, *12*(10), 10–15. https://doi.org/10.3390/app12104859

Muszyński, H. (1972). *Ideał i cele wychowania [The ideal and goals of upbringing]*. Biblioteka Nauczyciela PZWS.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nathiya, G., Punitha, S. C., & Punithavalli, M. (2010). *An Analytical Study on Behavior of Clusters Using K Means, EM and K* Means Algorithm*, *7*(3), 185–190. http://arxiv.org/abs/1004.1743

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press (Elsevier).

Nowak, S. (2007). *Metodologia badań społecznych*. Wydawnictwo Naukowe PWN.

Ocklind, C. (2023). A Comparative Analysis of Multilayer Network Software Bachelor's Programme in Computer Science. In *Bachelor's Programme in Computer Science* (Issue June). Uppsala Universitet.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report*, *4*, 1–100. http://arxiv.org/abs/2303.08774

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., … Malkov, Y. (2024). *GPT-4o System Card*, 1–33. http://arxiv.org/abs/2410.21276

Osowski, S. (1994). *Sieci neuronowe*. Oficyna Wydawnicza Politechniki Warszawskiej.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Plis, S. M., Hjelm, D. R., Slakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., Johnson, H., Paulsen, J., Turner, J., & Calhoun, V. D. (2014). Deep learning for neuroimaging: A validation study. *Frontiers in Neuroscience*, *8*, 1–11. https://doi.org/10.3389/fnins.2014.00229

Ptaszek, G. (2019). *Edukacja medialna 3.0: krytyczne rozumienie mediów cyfrowych w dobie Big Data i algorytmizacji*. Wydawnictwo Uniwersytetu Jagiellońskiego.

Puertas, E., Moreno-Sandoval, L. G., Redondo, J., Alvarado-Valencia, J. A., & Pomares-Quimbaya, A. (2021). Detection of Sociolinguistic Features in Digital Social Networks for the Detection of Communities. *Cognitive Computation*, *13*(2), 518–537. https://doi.org/10.1007/s12559-021-09818-9

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Retta, E. A., Sutcliffe, R., Mahmood, J., Berwo, M. A., Almekhlafi, E., Khan, S. A., Chaudhry, S. A., Mhamed, M., & Feng, J. (2023). Cross-Corpus Multilingual Speech Emotion Recognition: Amharic vs. Other Languages. *Applied Sciences (Switzerland)*, *13*(23). https://doi.org/10.3390/app132312587

Reykowski, J. (1966). *Funkcjonowanie osobowości w warunkach stresu psychologicznego [Personality functioning under psychological stress]*. Wydawnictwo Naukowe PWN.

Ríssola, E. A. (2020). *Text Mining for Online Mental State and Personality Assessment* (Issue September). Università della Svizzera Italiana.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rosseel, Y. (2012). *lavaan: An R Package for Structural Equation Modeling.* Journal of Statistical Software, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ruder, S., Peters, M., Swayamdipta, S., & Wolf, T. (2019). Transfer learning in natural language processing tutorial. *NAACL HLT 2019–2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Tutorial Abstracts*, 15–18.

Rudžionis, V., Lopata, A., Gudas, S., Butleris, R., Veitaitė, I., Dilijonas, D., Grišius, E., Zwitserloot, M., & Rudzioniene, K. (2022). Identifying Irregular Financial Operations Using

Accountant Comments and Natural Language Processing Techniques. *Applied Sciences (Switzerland)*, *12*(17). https://doi.org/10.3390/app12178558

Rutkowski, L. (2006). *Metody i techniki sztucznej inteligencji [Methods and Techniques of Artificial Intelligence]*. Wydawnictwo Naukowe PWN.

Rymanowski, M. (2007). *Modelowanie matematyczne i badanie złożonych układów analitycznych*. https://repozytorium.biblos.pk.edu.pl/redo/resources/26585/file/suwFiles/RymanowskiM_ModelowanieMatematyczne.pdf

Rytel, J. (2021). Kontrowersje wokół pojęcia trafności. *Studia Psychologica: Theoria et Praxis*, *21*(1), 49–63. https://doi.org/10.21697/sp.2021.21.1.03

Rzechowska, E. (2002). TPD as a starting point for micro-dynamic diagnosing of human development: theoretical and methodological reflections. In N. Duda (Ed.), *Positive Disintegration. The Theory of the Future* (pp. 283–296). Institute for Positive Disintegration in Human Development.

Rzechowska, E. (2004). *Potencjalność w procesie rozwoju: mikroanaliza konstruowania wiedzy w dziecięcych interakcjach rówieśniczych*. Wydawnictwo KUL.

Rzechowska, E. (2011a). Developmental Transformations: the Feeling of Meaning among Women in Emerging Adulthood. In M. T. E. Rzechowska, S. Steuden, D. Musial, E. Rydz (Eds.), *Contemporary Interpretations and Applications of the Theory of Positive Disintegration* (pp. 43–75). TN KUL.

Rzechowska, E. (2011b). *Dojrzały pracownik na rynku pracy: jak zabezpieczyć przed wykluczeniem społecznym osoby 50+*. Lubelska Szkoła Biznesu.

Rzechowska, E. (2011c). Podejście procesualne: warianty badań nad procesami w mikro- i makroskali. *Roczniki Psychologiczne*, *14*(1), 127–157.

Rzechowska, E., & Szymańska, A. (2017). Wykorzystanie strategii Rekonstrukcji Transformacji Procesu do budowy skali psychologicznej. In W. J. Paluchowski (Ed.), *Diagnozowanie – wyzwania i konteksty* (pp. 31–58). Wydawnictwo Naukowe Wydziału Nauk Społecznych UAM.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*, 2–6. http://arxiv.org/abs/1910.01108

Sarno, R., Dewandono, R. D., Ahmad, T., Naufal, M. F., & Sinaga, F. (2015). Hybrid association rule learning and process mining for fraud detection. *IAENG International Journal of Computer Science*, *42*(2), 59–72.

Schaeffer, E. S. (1959). A circumplex model for maternal behavior. *Journal of Abnormal Social Psychology*, *59*, 226–235.

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J. E., Demirutku, K., Dirilen-Gumus, O., & Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, *103*(4), 663–688. https://doi.org/10.1037/a0029393

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M., & Lepri, B. (2017). *What your Facebook Profile Picture Reveals about your Personality*. https://doi.org/10.1145/3123266.3123331

Shi, Z. R., Wang, C., & Fang, F. (2020). *Artificial Intelligence for Social Good: A Survey*. 1–78. http://arxiv.org/abs/2001.01818

Shin, H. (2022). Social contagion of academic behavior: Comparing social networks of close friends and admired peers. *PLoS ONE*, *17*(3 March), 1–16. https://doi.org/10.1371/journal.pone.0265385

Sierocki, R. (2020). Analiza sieci społecznych jako metoda badawcza w naukach społecznych. *Rocznik Antropologii Historii*, *X*(13), 223–255. https://doi.org/10.25945/rah2020.13.009

Sokołowski A., & Kosmol, J. (1995). Diagnostyka narzędzia za pomocą sieci neuronowej. In *Prace Naukowe Instytutu Technologii Maszyn i Automatyzacji Politechniki Wrocławskiej*. Oficyna Wydawnicza Politechniki Wrocławskiej.

Sośnicki, K. (1966). *Istota i cele wychowania [The essence and goals of upbringing]*. Nasza Księgarnia.

Srikant, R., & Agrawal, R. (1995). *Mining generalized association rules* (IBM Research Report No. RJ9963). IBM Almaden Research Center.

Srinadh, V. (2022). Evaluation of Apriori, FP growth and Eclat association rule mining algorithms. *International Journal of Health Sciences*, *6*(April), 7475–7485. https://doi.org/10.53730/ijhs.v6ns2.6729

STATISTICA Electronic Manual. (2012). *StatSoft Inc.* Pozyskane z http://documentation.statsoft.com/ STATISTICAHelp.aspx?path=common/AboutSTATISTICA/ElectronicManualIndex.

Steinberg, D. (2015). *CART: Classification and Regression Trees*. *December*.

Stephenson, D. (2018). *BIG DATA, nauka o danych i AI bez tajemnic*. HELION.

Stoica, A. A., Litvak, N., & Chaintreau, A. (2024). Fairness Rising from the Ranks: HITS and PageRank on Homophilic Networks. *WWW 2024 – Proceedings of the ACM Web Conference*, 2594–2602. https://doi.org/10.1145/3589334.3645609

Strus, W., Cieciuch, J., & Rowiński, T. (2014). The Circumplex of Personality Metatraits: A Synthesizing Model of Personality Based on the Big Five. *Review of General Psychology*, *18*(4), 273–286. https://doi.org/10.1037/gpr0000017

Suppes, P. (1972). *Axiomatic Set Theory*. Dover Publications.

Svenson, P. (2006). Complex networks and social network analysis in information fusion. *2006 9th International Conference on Information Fusion, FUSION*, *June*. https://doi.org/10.1109/ICIF.2006.301554

Szczęsny, W., Ciok, A., Kowalczyk, T., Pleszczyńska, E., & Wysocki, W. (1998). Gradacyjna analiza powiązań w tablicach kontyngencji. Zastosowania do analizy wyników głosowania do Sejmu RP z lat 1993 i 1997. *Studia Socjologiczne*, *3*(150).

Szczygieł, M. (2025). *Zaburzenia osobowości a skłonności do seryjnych morderstw. Analiza przypadków polskich morderców* (praca magisterska, Uniwersytet Kardynała Stefana Wyszyńskiego w Warszawie, promotor: dr hab. Agnieszka Szymańska, prof. UKSW).

Szymańska, A. (2009). Aspekty kontroli rodzicielskiej i dyscyplinowania dziecka [Aspects of parental control and disciplining of children]. *Psychologia Rozwojowa*, *14*(1), 37–47.

Szymańska, A. (2012). Parental Directiveness as a Predictor of Children's Behavior at Kindergarten. *Psychology of Language and Communication*, *16*(3), 1–24.

Szymańska, A. (2016a). Problematyka hierarchiczności – wyprowadzenie metacech w modelach SEM. In *Diagnoza psychologiczna jako przedmiot badania i nauczania*, Poznań.

Szymańska, A. (2016b). Założenia formalne modeli weryfikowanych przy pomocy układów równań strukturalnych. *Studia Psychologica. Theoria et Praxis*, *16*(2), 93–115.

Szymańska, A. (2017a). Problematyka hierarchiczności – wyprowadzanie metacech w modelach SEM [The issue of hierarchical models – the construction of meta-features in structural equation models]. *Studia Psychologica. Theoria et Praxis*, *1*, 65–84.

Szymańska, A. (2017b). Wykorzystanie algorytmów Text Mining do analizy danych tekstowych w psychologii [Usage of text mining algorithms to analyze textual data in psychology]. *Socjolingwistyka*, *33*, 99–116.

Szymańska, A. (2017c). Wykorzystanie analizy skupień metodą data mining do wykreślania profili osób badanych. *Studia Psychologiczne*, *55*, 26–42. https://doi.org/10.2478/V1067-010-0160-1

Szymańska, A. (2017d). Wykorzystanie analizy skupień metodą data mining do wykreślania profili osób badanych w badaniach psychologicznych [Using cluster analysis in the data mining method to draw profiles of participants surveyed in psychological research]. *Studia Psychologiczne*, *55*(1), 25–40.

Szymańska, A. (2018). Predicting model for aggressive directiveness in the light of Tadeusz Tomaszewski's theory of action: structural and data mining approach. *Psychology of Language and Communication*, *22*(1), 354–371. https://doi.org/10.2478/plc-2018-0016

Szymańska, A. (2019). *The transfer of parental mistakes in the family of origin of mothers of pre-school children: A structural and artificial intelligence approach*. Wydawnictwo Naukowe Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie.

Szymańska, A. (2023a). Funkcjonowanie osobowości depresyjnej w ujęciu psychocybernetycznym. *Zagadnienia Społeczne*, *1*(18), 114–131. https://nwsp.bialystok.pl/nr-1-1-2014/

Szymańska, A. (2023b). Przemoc wobec dziecka: błędy agresji, obojętności oraz ulegania a kształtowanie się osobowości antyspołecznej u kobiet. *Studia z Teorii Wychowania*, *42*(1), 147–164. https://doi.org/10.5604/01.3001.0016.3430

Szymańska, A. (2024a). Mechanizm identyfikacji projekcyjnej w interakcji wychowawczej: Perspektywa cybernetyki. *32. Ogólnopolska Konferencja Psychologii Rozwojowej – „Wspieranie Rozwoju – Postęp i Przemiana"*.

Szymańska, A. (2024b). Sztuczna inteligencja i systemy ekspertowe w diagnozie i pomocy psychologicznej i psychoterapeutycznej. In *Sztuczna inteligencja – szanse i zagrożenia* (pp. 1–305). Wydawnictwo Naukowe Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie.

Szymańska, A. (2024c). Zastosowanie analizy sieci społecznej i drzewa decyzyjnego w badaniu relacji między zaburzeniami osobowości. *Zagadnienia Społeczne*, *1*(19).

Szymańska, A. (2025a). Formal Assumptions and Limitations of Circular Models in Typologizing Psychological and Educational Theories. *Studia z Teorii Wychowania*, *2*(15), 97–117.

Szymańska, A. (2025b). *Profiling psychological traits in kernel space: From the 2D plane to the Reproducing Kernel Hilbert Space (RKHS)* [Preprint]. PsyArXiv. Retrieved from https://osf.io/bzucx/

Szymańska, A. (2025c). Psychometric Models in Higher Dimensions: How Artificial Intelligence Can Expand the Space of Measurement. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/pe67q_v2

Szymańska, A. (2025d). Topologization in Psychological Modeling: From Two-Dimensional Analysis to the Third Dimension in Psychometrics. *PsyArXiv*. https://doi.org/10.31234/osf.io/fd4sb_v1

Szymańska, A., & Aranowska, E. (2016). *Błąd w wychowaniu. W stronę weryfikacji teorii Antoniny Guryckiej [Mistake in upbringing. Towards a verification of Antonina Guryc-ka's theory]*. Liberi Libri.

Szymańska, A., & Aranowska, E. (2019). Parental Stress in the Relationship with the Child and Personality Traits that Parents Shape in their Children. *Early Child Development and Care*. https://doi.org/10.1080/03004430.2019.1611569

Szymańska, A., & Aranowska, E. (2022). Raising a child to live in society – Personality traits parents develop and prevent from developing in their preschool children. *Studia z Teorii Wychowania*, *4*(41), 409–431. https://doi.org/DOI: 10.5604/01.3001.0016.1654

Szymańska, A., & Aranowska, E. (2023). Raising a Boy and a Girl: Personality Traits that Mothers Develop and Prevent from Developing in their Preschool-age Sons and Daughters. *Forum Pedagogiczne*, *13*(1), 335–351. https://doi.org/https://doi.org/10.21697/fp.2023.1.23

Szymańska, A., Grzesiuk, L., Suszek, H., Dobrenko, K., Krawczyk, K., & Rutkowska, M. (2018). Badania polskich psychoterapeutów – z jakimi pacjentami pracują i jakie stosują metody psychoterapii [Research on Polish psychotherapists – what types of patients they work with and what methods of psychotherapy they use]. *Psychiatria Polska, 52*(4), 749–769. https://doi.org/10.12740/PP/OnlineFirst/70462

Szymańska, A., & Torebko, K. (2015). Struktura błędu wychowawczego. Weryfikacja struktury zaproponowanej w modelu kołowym przez Antoninę Gurycką. *Studia z Teorii Wychowania*, *3*(12), 165–192.

Śliwińska, M., Zawadzki, B., & Strelau, J. (1995). Adaptacja "Zmodyfikowanego kwestionariusza wymiarów temperamentu" Windle'a iLernera do warunków polskich: zastosowanie do diagnozy temperamentu młodzieży i osób dorosłych. *Studia Psychologiczne*, *1–2*, 113–146.

Tadeusiewicz, R. (1993). *Sieci neuronowe*. Akademicka Oficyna Wydawnicza.

Tadeusiewicz, R. (2001). *Wprowadzenie do sieci neuronowych*. StatSoft Polska.

Tadeusiewicz, R. (2012). Komputerowe wspomaganie decyzji. *XVI Konferencja Automatyków*, 1–17.

Tadeusiewicz, R., Gąciarz, T., Borowik, B., & Leper, B. (2007). *Odkrywanie właściwości sieci neuronowych*. Polska Akademia Umiejętności.

Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text Mining: Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, *7*(11), 414–418. https://doi.org/10.14569/ijacsa.2016.071153

Tang, X. (2024). A latent hidden Markov model for process data. *Psychometrika*, *89*(1), 205–240.

Tarwacka-Odolczyk, A., Tomaszewski, P., Szymańska, A., & Bokus, B. (2014). Deaf children building narrative texts. Effect of adult-shared vs. non-shared perception of a picture story. *Psychology of Language and Communication*, *18*(2), 149–177.

Tomaszewski, T. (1975). Człowiek w sytuacji. In T. Tomaszewski (Ed.), *Psychologia* (pp. 17–36). Państwowe Wydawnictwo Naukowe.

Tomaszewski, T. (1982). *Psychologia.* Państwowe Wydawnictwo Naukowe.

Turner, J. S., & Helms, D. B. (1999). *Rozwój człowieka*. Wydawnictwa Szkolne i Pedagogiczne.

Twisk, Jos W. R. (2010). *Analiza wielopoziomowa – przykłady zastosowań. Praktyczny podręcznik biostatystyki i epidemiologii*. Wydawnictwo Szkoły Głównej Handlowej w Warszawie.

Umami, M. H., Prihandini, R. M., & Agatha, A. B. (2024). *Application of graph theory to social network analysis* [Unpublished manuscript]. Department of Mathematics Education, University of Jember, Indonesia.

Vandeleest, J. J., Beisner, B. A., Hannibal, D. L., Nathman, A. C., Capitanio, J. P., Hsieh, F., Atwill, E. R., & McCowan, B. (2016). Decoupling social status and status certainty effects on health in macaques: A network approach. *PeerJ*, *4*, e2394. https://doi.org/10.7717/peerj.2394

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All You Need*. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates. Retrieved from https://papers.nips.cc/paper/7181-attention-is-all-you-need

Verma, M., & Mehta, D. (2014). Sequential Pattern Mining: A Comparison between GSP, SPADE and Prefix SPAN. *International Journal of Engineering Development and Research*, *2*(3), 2321–9939. www.ijedr.org

Wang, L. (2005). *Support Vector Machines: Theory and Applications*. Springer. https://books.google.com/books?hl=en&lr=&id=uTzMPJjVjsMC&oi=fnd&pg=PA1&dq=support+vector+machines&ots=GFAK9w2Hfb&sig=4AddZM1BrpsopEIiErlIzeys6zI

Wang, W., Kofler, L., Lindgren, C., Lobel, M., Murphy, A., Tong, Q., & Pickering, K. (2023). AI for Psychometrics: Validating Machine Learning Models in Measuring Emotional Intelligence with Eye-Tracking Techniques. *Journal of Intelligence*, *11*(9). https://doi.org/10.3390/jintelligence11090170

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 1–47.

Warchalska-Troll, A., & Warchalski, T. (2022). The selection of areas for case study research in socio-economic geography with the application of *k*-means clustering. *Wiadomości Statystyczne. The Polish Statistician*, *67*(2), 1–20. https://doi.org/10.5604/01.3001.0015.7717

Ważyńska, A., Szymańska, A., Bartczak, M., & Bokus, B. (2015). Przy okrągłym stole dialogowego ja. Gdzie siedzi sceptyk? In B. Bokus & E. Kosowska (Eds.), *O wątpieniu* (pp. 63–82). Studio Lexem.

Wikipedia contributors. (2024, January 30). *ELIZA*. In *Wikipedia*. https://en.wikipedia.org/wiki/ELIZA

Wójtowicz, A. (1989). Błąd wychowawczy w doświadczeniach młodzieży. In A. Gurycka & A. Gołąb (Eds.), *Podmiotowość w doświadczeniach wychowawczych dzieci i młodzieży. Wychowanek jako podmiot doświadczeń* (pp. 81–106). Wydawnictwa Uniwersytetu Warszawskiego.

Wołowiec-Korecka, E. (2016). Zastosowanie sztucznych sieci neuronowych do modelowania procesów azotowania próżniowego stali narzędziowych. In *Zastosowania statystyki i data mining w badaniach naukowych* (pp. 21–36). StatSoft Polska.

Wright, S. A., & Schultz, A. E. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons*, *61*(6), 823–832.

Yenduri, G., Ramalingam, M., Selvi, G.C., Supriya, Y., Srivastava, G., Maddikunta, P.K.R., Raj, G.D., Jhaveri, R.H., Prabadevi, B., Wang, W., Vasilakos, A.V., & Gadekallu, T.R. (2024). GPT (Generative Pre-Trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *IEEE Access*, *12*(March), 54608–54649. https://doi.org/10.1109/ACCESS.2024.3389497

Yim, H., Boo, Y., & Ebbeck, M. (2014). A Study of Children's Musical Preference: A Data Mining Approach. *Australian Journal of Teacher Education*, *39*(2), 21–34.

Youn, G., Yoon, B., Ji, S., Ko, D., & Rhee, J. (2022). Mixup Based Cross-Consistency Training for Named Entity Recognition. *Applied Sciences (Switzerland)*, *12*(21). https://doi.org/10.3390/app122111084

Yuan, S. (2023). Design and Visualization of Python Web Scraping Based on Third-Party Libraries and Selenium Tools. *Academic Journal of Computing & Information Science*, *6*(9), 25–31. https://doi.org/10.25236/ajcis.2023.060904

Zainol, Z., Wani, S., Nohuddin, P. N. E., Noormanshah, W. M. U., & Marzukhi, S. (2018). Association Analysis of Cyberbullying on Social Media using Apriori Algorithm. *International Journal of Engineering & Technology*, *7*(4.29), 72–75. https://doi.org/10.14419/ijet.v7i4.29.21847

Zhang, Q., & Segall, R. S. (2008). Web Mining: A survey of current research, techniques, and software. *International Journal of Information Technology and Decision Making*, *7*(4), 683–720. https://doi.org/10.1142/S0219622008003150

Żurada, J., Barski, M., & Jędruch, W. (1992). *Sztuczne sieci neuronowe*. Państwowe Wydawnictwo Naukowe.

APPENDICES

APPENDIX A

Research Procedure and Study Sample

The study was conducted online via the University Online Research System (UORS), where a set of questionnaires was made available regarding various aspects of parenting and parental personality. The research included, among others, questionnaires assessing *parental goals* in the family of origin, parental personality traits, value systems, psychological needs, perception of *parental mistakes* committed by grandparents, as well as surveys concerning the child—such as the child's age, gender, birth order, and number of siblings. Additional questionnaires addressed current parental goals, parenting difficulties (parental stress), stress responses, self-perception of parenting errors, and the child's temperamental traits.

Preschools were randomly selected from lists provided by Boards of Education in each province. The management of these preschools was informed about the opportunity to participate in the study and then relayed the information to parents via email or posters displayed on bulletin boards. Before starting the survey, parents were informed about the conditions of participation, the duration of the study, and the procedure for completing the questionnaires. They were also asked to focus on one selected preschool-aged child to avoid confusion or inaccuracies that could arise from answering on behalf of more than one child. The average time to complete the questionnaires was 70 minutes, and the online system monitored the response duration.

Initially, 2,183 individuals participated in the study, of whom 546 completed the entire protocol. The final sample consisted of 420 parents of preschool-aged children, including 402 mothers and 18 fathers. Due to the non-representative number of fathers, all analyses were performed exclusively on the group of mothers. The

size of this sample and the random selection of participants allow for generalization of the results to the population of mothers of preschool-aged children.

The participating mothers ranged in age from 21 to 50 years, with the largest proportion in the 28–39 age bracket. The most frequent age was 34 years, and the median was 33 years. Most mothers had higher education (84.8%), 14.7% had secondary education, and 0.4% had completed only primary or vocational education.

In terms of place of residence, 43.8% of the mothers lived in large cities (over 200,000 residents), 37.1% in smaller towns, and 19.2% in rural areas. The sample included a nearly equal number of mothers referring to girls and boys—50.5% of responses concerned girls, and 49.5% boys. The children's ages ranged from 3 to 6 years: 28.1% of the children were 3 years old, 24.1% were 4 years old, 20.9% were 5 years old, and 26.9% were 6 years old.

The gender distribution within each age group was balanced. Among 3-year-olds, 54.9% were boys and 45.1% girls; among 4-year-olds, 44.3% were boys and 55.7% girls; among 5-year-olds, 45.2% were boys and 54.8% girls; and among 6-year-olds, 51.9% were boys and 48.1% girls.

Regarding educational setting, 82.6% of the children attended preschool, with 47.3% enrolled in public preschools, 26.4% in private institutions, and the remainder in Catholic or Montessori schools. A small proportion of children were homeschooled by parents, grandparents, or a governess, and 5% of the children had already begun primary school.

Nearly 36% of the children were only children, 48% had one sibling, and the remaining 16% had more than one sibling. In terms of birth order, 82.1% of the children referred to by mothers were first-borns, 12.2% were second-born, and 5.8% were third or later-born within the family.

APPENDIX B

Research Procedure and Sample

The study was conducted between January and May 2007 in several preschools located in different towns: five preschools in Legionowo, one in Zegrze, three in Warsaw, three in Częstochowa, and one in Kraków. Teachers from 13 groups of six-year-olds, 13 groups of five-year-olds, and 10 groups of four-year-olds were asked to select three "difficult" children and three "well-behaved" children from each group, based on their common-sense understanding of these terms. This method of selection was intended to minimize the impact of individual differences in the interpretation of the term "parenting difficulties" among the teachers (in accordance with the methodology described by Dryll, 1995). Once the children had been selected, their parents were invited to participate in the study.

The research sample consisted of 102 women and 102 men, with an average participant age of 35 years. Among the women, the average age was 34, while among the men it was 35. In terms of education, 7 men and 4 women had vocational education, 32 men and 40 women had completed secondary education, and 49 men and 64 women held higher education degrees.

Parents were given a set of questionnaires to complete at home. This set included the DAiS questionnaire, a tool for measuring acceptance of control, and a questionnaire assessing parenting difficulties. Additionally, participants completed a shortened version of the Social Approval Scale (KAS), consisting of 13 items characterized by the highest reliability.

After collecting the questionnaires, the scores on the KAS scale were calculated, and any questionnaires in which the respondent scored more than 9 points out of a possible 13 were excluded from analysis. In addition, incomplete questionnaires

were also rejected. Ultimately, out of 262 completed questionnaires, 204 were qualified for analysis. This group included 51 questionnaires from mothers of "well-behaved" children, 51 from mothers of "difficult" children, 51 from fathers of "well-behaved" children, and 51 from fathers of "difficult" children.

APPENDIX C

Research Procedure and Sample

The study was conducted online across Poland, where parents were asked to complete previously prepared questionnaires regarding their children attending preschool. At the beginning of the study, parents were instructed to focus on one selected child to avoid response errors in cases where they had more than one child.

A total of 319 parents participated in the study, including both fathers and mothers of preschool-aged children. The participants were aged between 22 and 54 years, with the majority falling within the 28–35 age range, indicating that this was a group of young adults. The most frequent age among participants was 33, while the median age was 27.

In terms of education, most respondents held a higher education degree (63.4%), while a smaller proportion had secondary education (29.1%). A small group had completed primary or vocational education (4.6%), and only 2.9% of participants held a doctoral degree. Regarding place of residence, the largest proportion of respondents came from large cities (with over 500,000 inhabitants – 26.4%), while 22.3% lived in medium-sized cities (50,000–200,000 inhabitants). Individuals from small towns (up to 10,000 inhabitants) accounted for 6.8%, and 13.1% of respondents lived in rural areas.

Both mothers and fathers of boys and girls participated in the study, with a slight predominance of girls (55.5% compared to 44.5% boys). The children included in the study ranged from 3 to 6 years old, with 47.3% being 3–4 years old and 52.7% being 5–6 years old. Among the younger children (ages 3–4), the gender distribution was fairly balanced, while in the older age groups (5–6), there were slightly more girls than boys.

The majority of children attended public preschools (58.3%) or private ones (21%), while a smaller group was enrolled in other types of preschools, including Catholic and Montessori institutions. The latter, due to their relatively small number in the country, were the least represented in the study.

Agnieszka Szymańska

Mathematical Modeling in Psychology Using Artificial Intelligence

Reviewer: Prof. Ph. D. Czesław S. Nosal

English translation and proofreading: GPT-4 language model (Generative Pre-trained Transformer 4) by OpenAI

Cover design: Sora generative model by OpenAI

Proofreading: UKSW University Press editorial team

Typesetting and layout: Cyprian Pietrykowski

ISBN: 978-83-8281-641-9 (e-book)

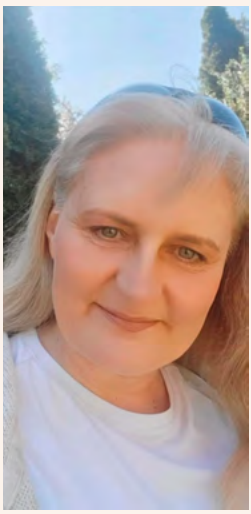
Wydawnictwo Naukowe UKSW

01-815 Warszawa, ul. Dewajtis 5

Phone: +48 22 561-89-23

e-mail: wydawnictwo@uksw.edu.pl

www.wydawnictwo.uksw.edu.pl



About the Author

Dr hab. Agnieszka Szymańska, DSc, Associate Professor at UKSW – holds a Doctor of Social Sciences degree (DSc, dr hab.) in the discipline of psychology and serves as an associate professor in the Department of Psychological Research Methodology at the Institute of Psychology, Cardinal Stefan Wyszyński University in Warsaw. She specializes in psychological research methodology, statistics, psychometrics, and the applications of artificial intelligence in psychology.

She completed postgraduate studies in mathematics, and she also completed a specialization in psychometrics at the Faculty of Psychology of the University of Warsaw. She lectures on artificial intelligence in the Big Data in Social Analytics program at UKSW and, for many years, taught courses in this area within the university's doctoral program.

She is a psychologist with a specialization in educational psychology and has also completed postgraduate training in psychotherapy.

This book is very important and necessary for doctoral students and psychology students, as well as for the broader community of researchers studying the multivariate determinants of behavior. It presents the formal foundations of many different structural analysis methods, both classical and new, within the computational procedures of artificial intelligence. The author of the book accurately, and without exaggeration, recognizes the role of artificial intelligence frameworks in certain statistical procedures, while at the same time emphasizing the importance of interpreting the results obtained by researchers.

Czesław S. Nosal