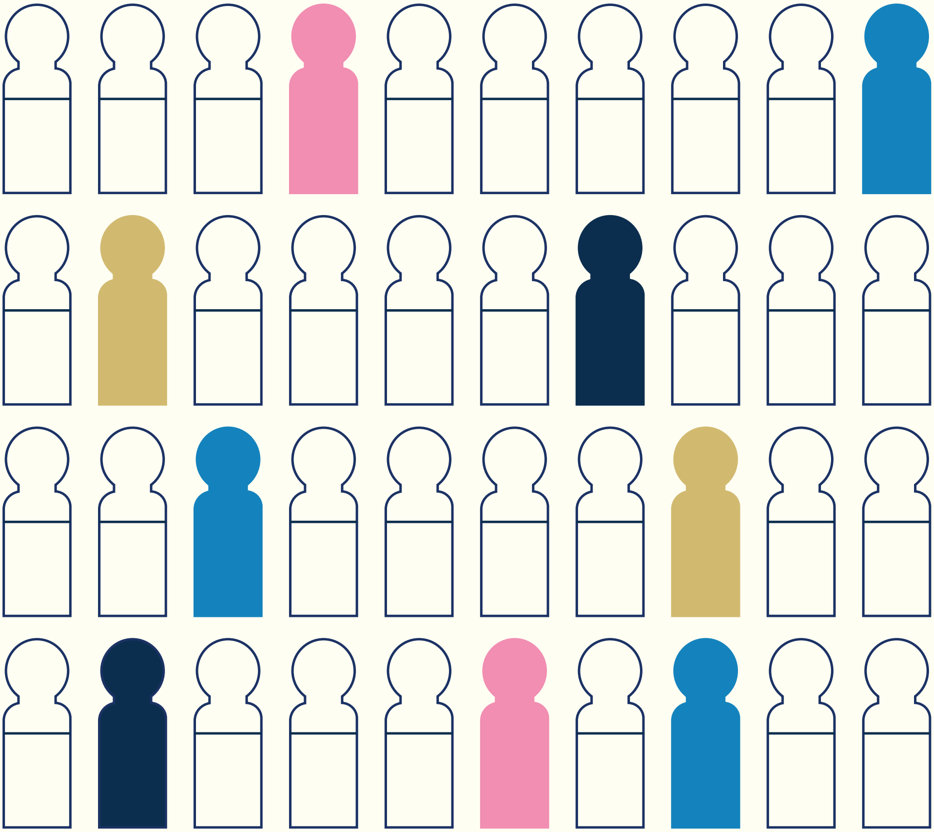
Exceptional Model Mining for Hierarchical Data

Rianne Margaretha Schouten



Exceptional Model Mining for Hierarchical Data

Rianne Margaretha Schouten

Colophon

SIKS Dissertation Series No. 2025-04
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Printed by: Grefo Prepress

Cover design: Maaike Schouten (https://artbymaaike.nl)

A catalogue record is available from the Eindhoven University of Technology Library with ISBN 978-90-386-6238-1, and an electronic version of this dissertation is available online at https://research.tue.nl.

Exceptional Model Mining for Hierarchical Data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van
de rector magnificus prof. dr. S.K. Lenaerts,
voor een commissie aangewezen door het College
voor Promoties, in het openbaar te verdedigen op
donderdag 16 januari 2025 om 13:30 uur

door

Rianne Margaretha Schouten

geboren te Goor, Nederland

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

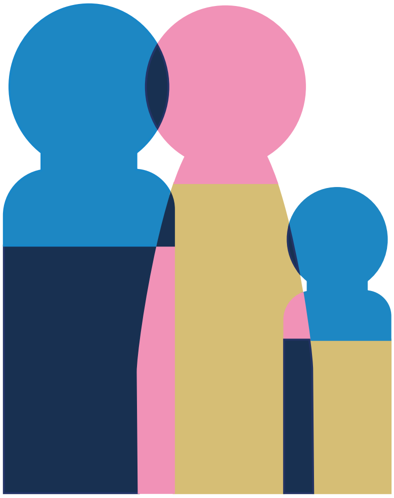
voorzitter:	prof. dr. F.C.R. Spieksma
promotor:	prof. dr. M. Pechenizkiy
co-promotor:	dr. W. Duivesteijn
leden:	prof. dr. M. Atzmueller (Osnabrück University)
	prof. dr. B. Hammer (Bielefeld University)
	prof. dr. C. Plant (University of Vienna)
	prof. dr. B. Crémilleux (University of Caen Normandy)
	prof. dr. A. Vilanova

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

If your dreams do not scare you, they are not big enough.

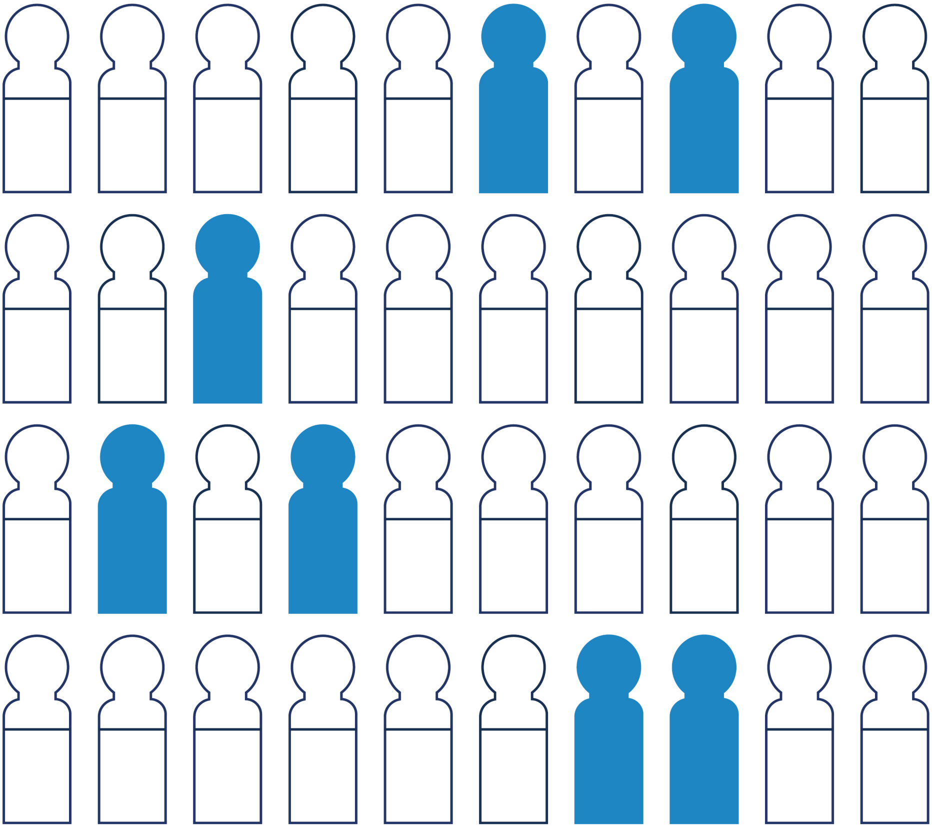
Ellen Johnson Sirleaf
Muhammad Ali
Sifan Hassan

To my two men



Contents

1



Introduction

In this chapter, we introduce the topic of this dissertation. We are motivated to contribute to further understanding variation in human behavior by extracting relevant and meaningful patterns from data using a Local Pattern Mining framework called Exceptional Model Mining (EMM). We observe that employing EMM in the real world is challenging, since real-world data is often hierarchically structured. Therefore, the main research question in this dissertation is: How to discover exceptional subgroups in hierarchical data? The work presented in this dissertation significantly contributes to the body of scientific literature on EMM for hierarchical data and presents new EMM methodologies for sequential, repeated cross-sectional and nested data.

1.1 Motivation

People differ, not only in physical appearance but in personality, cultural background, abilities and interests, as well as in cognitive, emotional and social behavior [44, 79, 145, 146]. Understanding variation in human behavior is valuable in many situations. Consider the domain of diabetes care, where e-coaching tools support patients in the self-management of their chronic condition. Here, it is key to provide feedback and recommendations to people personally, based on individual characteristics.

In this dissertation, we analyze variation in human behavior using a Local Pattern Mining (LPM) framework called Exceptional Model Mining (EMM) [54, 121]. LPM is a subdomain of Data Mining (DM), the process of extracting patterns from data [58], and aims to discover *local* patterns. In contrast to *global* patterns that explain most of the instances in the data, local patterns cover small parts of the data space, deviate from the distribution of the population and show some internal structure [74, 139].

The framework of EMM focuses on discovering subgroups in a population that somehow behave exceptionally [54, 121]. It describes these subgroups using a rule-based language based on conjunctions of selection conditions on a pre-defined subset of data attributes, the descriptors. Exceptional behavior is measured in terms of parameters of a model class over other attributes, the targets. EMM is particularly valuable for extracting patterns that are practically useful. For instance, in Section 5.5, we discover patients with exceptional blood glucose fluctuations. These patients are described by *HbA_{1c} category = low and diabetes duration ≥ 20 years and 21.3 ≥ BMI ≥ 35.7*. It is evident that discovering such an exceptional subgroup provides highly relevant information for clinicians and practitioners about the needs of individual patients.

We observe that employing EMM in real-world use cases is challenging, since data is often hierarchically structured. In short, hierarchical data contains observations on more than one entity type, where the observations of one entity type are nested in the entities of another entity type. The concept is well-known in the social and biomedical sciences and stems from the idea that individual persons are influenced by the social groups or contexts to which they belong (and vice versa) [84, 125, 187]. For example, self-rated health may relate to the population density of neighborhoods [125] and students' performance may depend on the skills of their teacher [84]. The individuals and social groups are conceptualized as a hierarchical system of individuals nested in groups, and groups nested in larger groups. Then, a shared context introduces a correlation structure between individuals belonging to that context and a common data mining assumption that observations are independent is violated.

In hierarchical data, the lowest level is not necessarily that of the individual. For instance, in Section 8.7, students perform multiple tasks, with multiple items per task. This results in hierarchical data where item-level attribute information is nested in tasks, and tasks are nested in students. Another type of hierarchical data may occur when information is repeatedly measured. For instance, in Section 5.3, glycemic treatment is monitored by repeatedly sampling blood glucose values. These measurements are not independent but should be considered nested in (the context of) individuals.

Formally, definitions in EMM are agnostic about the origin of the data, whether or not one observation is independent from the next one and whether or not the data should be formatted as a flat table [54, 121]. However, most existing EMM methodologies assume data to be in the conventional data mining representation where each individual can be described with one tuple of attribute-values and where a column contains the same semantic information for each individual. In contrast, in hierarchical data, attributes contain multiple values per individual and individuals are described by tuples of tuples (i.e., a sequence or tuple of values per attribute). Then, the EMM framework runs into problems with the selection and description of candidate subgroups and with assessing exceptionality. Hence, further development is needed.

1.2 Objective

We aim to extract societal relevant patterns from data and to enable domain experts to answer important, real-world questions. However, real-world data often has a hierarchical structure, which poses problems for the framework of EMM. We therefore need to solve problems regarding whether and how hierarchical data can be formatted as a flat table, how selection conditions can be used to cover a group of individuals and how we can assess exceptionality. The main research question in this dissertation is:

How to discover exceptional subgroups in hierarchical data?

1.3 Research approach

To answer the main research question, in Section 3.4 we build on terminology from multilevel analysis (MLA) and databases and formally define hierarchical data as a collection of measurements taken from various types of entities, where the measurements on the entities of one entity type are nested in the entities of another entity type. For instance, consider observations from *students, nested in classes* or *repeated measurements, nested in individuals*. We then explore existing approaches for EMM for hierarchical data. Recall that in EMM, data attributes are divided into descriptors, used to create and describe subgroups, and targets, used to model and quantify exceptionality. We propose a unified framework for EMM for hierarchical data based on whether the hierarchy exists on the descriptive side, on the target side or on both sides.

We closely work together with domain experts to address important problems in the real world. To that end, we follow a research circuit as depicted in Figure 1.1, where we start in the left cycle by formulating *domain-inspired research questions* (D-RQs). Subsequently, we investigate the data structure in descriptive and target space and explore whether existing EMM methodology can be deployed to extract relevant patterns. For all D-RQs discussed in this dissertation, collected data was hierarchically structured and could not be analyzed using existing techniques.

We then traverse to the right cycle and formulate *EMM-related research questions* (EMM-RQs). We develop generic and domain-independent solutions, and validate our methods on synthetic and public data. Remark that we do more than just evaluating the internal validity of our methods: we circle back to the left cycle in Figure 1.1 and additionally assess external validity by demonstrating that our proposed EMM methodologies enable domain experts to confirm existing hypotheses and to spawn interest for new theories. In all chapters in this dissertation, domain experts interpreted our findings and confirmed that our discovered patterns truly contribute to answering domain-specific questions. Next, we give an overview of the D-RQs and EMM-RQs for three domains of interest: diabetes care, public health and learning analytics.

1.3.1 Diabetes care

The clinical accepted standard for monitoring glycemic treatment for patients with diabetes type 2 is to measure the blood level of glycated haemoglobin ($\mathrm{HbA_{1c}}$) [208]. However, the use of $\mathrm{HbA_{1c}}$ has important limitations: its assessment does not contribute to reduction of hypoglycemic episodes and it does not reflect blood glucose fluctuations [38, 108]. An alternative is to investigate the association between $\mathrm{HbA_{1c}}$ and blood glucose values measured with an iCGM device [46]. Consequently, [46, p. 2244] formulate the following domain-inspired research question:

D-RQ Diabetes care *How can the use of iCGM-derived parameters support establishing individualized glycemic treatment?*

In this dissertation, we aim to contribute to answering this question by developing an EMM methodology that discovers subgroups of patients with exceptional blood glucose fluctuations. We use data collected by the second DIAbetes and LifEstyle Cohort Twente (DIALECT-2) [46, 67]. In descriptive space, the DIALECT-2 dataset has the conventional data mining representation where each patient is represented by a tuple of attribute values. In target space, the iCGM measurements form sequences (or, time series) of blood glucose values, where each sequence belongs to an individual patient. In other words, observations in target space reside at a lower hierarchical level than observations in descriptive space.

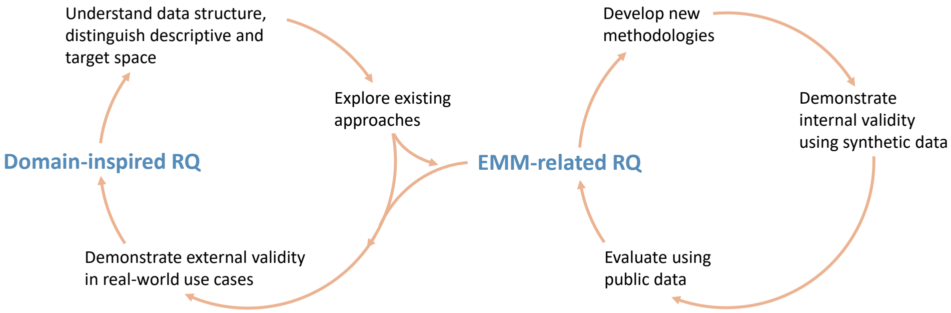

Understand data structure, distinguish descriptive and target space
Explore existing approaches
Domain-inspired RQ
Demonstrate external validity in real-world use cases
Develop new methodologies
Demonstrate internal validity using synthetic data
EMM-related RQ
Evaluate using public data

Figure 1.1: Schematic illustration of the research circuit used in this dissertation. We aim to contribute to answering domain-inspired research questions using the framework of Exceptional Model Mining (EMM). In EMM, data attributes are divided into a set of descriptors, used to create and describe subgroups, and a set of targets, used to model and quantify exceptionality. A hierarchical data structure may occur in descriptive space, in target space or in both. This brings about EMM-related research questions.

Although there exist global pattern mining methods for analyzing time series data with additional attribute information, in EMM such a combination is rare. An EMM model class exists for 1st order Markov chains [124]. However, [124] re-format the data by splitting sequences of length T into $T-1$ transitions. Instead, we aim to discover patients with exceptional blood glucose fluctuations and therefore require that an individual's sequence is selected into the subgroup in its entirety; it is not meaningful to split sequences. Consequently, compared to [124] who detect heterogeneity within sequences, we aim to discover subgroups of homogeneous sequences that are heterogeneous with respect to the overall population.

To effectively quantify exceptionality in sequential data, we aim for our target model to be applicable to sequences of varying lengths. Furthermore, our method should allow for the incorporation of target models that use a varying degree of memory. Therefore, the EMM-RQs are:

EMM-RQ A1 *Which type of target model could provide valuable parameters of blood glucose fluctuations?*

EMM-RQ A2 *How to handle varying sequence-lengths in target space?*

EMM-RQ A3 *How to assess exceptionality of subgroups when the number of model parameters differs between subgroups?*

1.3.2 Public health

Analyzing societal trends is an important line of research in social sciences as it assesses how the behaviors, attitudes, and feelings of populations change over periods of time, and for which groups such changes are particularly pronounced. Current strategies for analyzing adolescent alcohol use focus on the average, general trend in the population [161, 192]. For instance, the percentage of Dutch adolescents that had consumed alcohol in the last 4 weeks at the time of measurement has decreased from 57% in 2003 to 26% in 2015 and has flattened since then [161, 192]. It would be valuable to understand how and why the trend in alcohol use differs among subgroups of adolescents [39–41]. Such an understanding helps policy makers, government institutions and decision makers to take the right course of action. The D-RQ is formulated as follows:

D-RQ Public health *How can we obtain a better understanding of factors that influence when, whether, and why the downward trend in adolescent alcohol use in the Netherlands flat-lines?*

In the Netherlands, the Dutch National School Survey on Substance Use (DNSSSU) [161] and the Health Behavior in School-aged Children study (HBSC) [192] investigate alcohol use of Dutch adolescents. Both DNSSSU and HBSC adopt a Repeated Cross-Sectional (RCS) research design where new cases are repeatedly sampled from a population at successive measurement moments (both DNSSSU and HBSC are conducted every four years, with an offset of two years between them). Consequently, in RCS data, given the measurement occasion, cases are independent, but over the entire trend period, cases from the same measurement occasion are more alike than cases from separate occasions.

We aim to deploy the EMM framework to discover subgroups of adolescents with exceptional trend deviations. We cannot directly apply existing instances of EMM to RCS data for various reasons. First, no model class and quality measure exist that are suitable for analyzing trends. Compared to sequential or time series data where the sequence is known per individual and the sample size is fixed, in RCS data, an individual contributes to the trend at just one measurement occasion. Consequently, the entire trend is estimated on data with varying sample sizes. Second, in descriptive space, the distributions of socio-demographic variables change over time (e.g., the number of adolescents with a non-native background fluctuates [40]) and varying survey questions may alter the data type or induce missing values. Consequently, the EMM-RQs are:

EMM-RQ B1 *How to create and describe subgroups when descriptive attributes are incomplete?*

EMM-RQ B2 *Which problems occur when distributions of descriptors change over time and how can we mitigate these problems?*

EMM-RQ B3 *How to define and quantify exceptional deviations of societal trends?*

1.3.3 Learning analytics

Variation in number processing skills at a young age may predict achievements in mathematical skills later in life [71, 130]. Therefore, understanding individual variation in learning behavior at an early stage is crucial for providing targeted support to the needs of individual children, and for developing assessment tools and intervention programs that adapt to these individual needs [18, 155].

Specifically, enumeration performance reflects two distinct neural processes: the subitizing system where small sets (1-4 dots) are recognized accurately and rapidly, and the counting system where larger sets are enumerated more slowly, perhaps by counting dots or other enumeration strategies [157]. Individual differences in subitizing range predict math ability [130]. Therefore, the D-RQ is:

D-RQ Learning analytics *What are characteristics of subgroups of children whose subitizing curves exhibit atypical patterns?*

The FUnctional Numerical Assessment (FUNA) project [155] is a large-scale research program that develops digital assessment tools for detecting dyscalculia and dyslexia. In FUNA, numerical processing competences and math abilities are assessed using several computer-assisted tasks. Some of these tasks contain a fixed number or questions, or items; others are time-based and the number of answered items will vary per child. Items are taken from a larger set of items, and are not necessarily answered in the same order. Consequently, the dataset does not follow the conventional data mining representation where each individual can be described with one tuple of attribute-values. Rather, the information is presented as one tuple per attribute per child (thus, a tuple of tuples) and has as a hierarchical structure in both descriptive and target space.

Few existing literature considers lower-level descriptive attributes; most of them deploy Subgroup Discovery (SD) rather than EMM: they assume one categorical target attribute and use quality measures related to Weighted Relative Accuracy (WRAcc) [103, 209, 211] (more on the difference between EMM and SD can be found in Section 2.2). To the best of our knowledge, only two EMM methodologies have been developed for data with a nested structure in both descriptive and target space: [93] construct attributes from time series and [88] adopt an alternative description language. Instead, we aim to develop an instance of the EMM framework that answers the following questions:

EMM-RQ C1 *Which methods exist for flattening hierarchical descriptors, and what are their requirements, advantages and disadvantages?*

EMM-RQ C2 *How to formally define the approach where nested descriptive data is flattened using domain-specific aggregation functions?*

In target space, subitizing curves are estimated with segmented linear regression, but it is unknown which regression parameter may display exceptional behavior. In addition, the number of model parameters may differ between subgroups. Therefore, we formulate the research question as follows:

EMM-RQ C3 *How to quantify exceptionality with segmented linear regression as a target model, taking into account that the parameter displaying exceptional behavior is unknown and that the number of model parameters may differ between subgroups?*

1.4 Outline and contributions

Figure 1.2 gives a schematic overview of the structure of this dissertation. You are currently reading **Chapter 1**. In **Chapter 2**, we provide background information on how EMM originated as an LPM framework. The chapter also introduces traditional notation and definitions for EMM, and gives an overview of existing methods for reducing subgroup set redundancy and validating discovered subgroups. In this dissertation, we will apply several combinations of these anti-redundancy and validation techniques and sometimes experimentally evaluate their effects.

The main contributions of this dissertation can be found in Chapters 3 - 8. In **Chapter 3**, we provide a unified terminology for EMM for hierarchical data. We propose the notion of subgroup level and distinguish between descriptive and target attributes that abide at a lower, the same or a higher hierarchical level. We then categorize existing EMM approaches and uncover existing research gaps. Some of these will be filled by work proposed in this dissertation. In **Section 9.1**, we discuss how the respective chapters of this dissertation fit in the unified terminology.

In **Chapter 4**, we solve problems with analyzing sequential data in target space. We answer **EMM-RQs A1, A2 and A3** by proposing discrete Markov chains for modeling fluctuations in sequential data and quantify subgroup exceptionality using a log likelihood based quality measure, utilizing information-theoretic scoring functions such as Akaike's Information Criterion (AIC) [3, 4] and the Bayesian Information Criterion (BIC) [180]. These scoring functions allow us to fit Markov chains of varying order and to consider subgroup models that have a different number of parameters than the model fitted to the entire dataset. We demonstrate the effectiveness of our proposed approach through extensive synthetic data experiments and a public dataset.

We demonstrate external validity of our approach in **Chapter 5**. We contribute to answering **D-RQ Diabetes care** by discovering subgroups of patients with exceptional fluctuations in blood glucose values, based on patient-specific information such as age and HbA$_{1c}$ level. Clinicians and domain experts confirmed the blood glucose transition behavior as estimated by the fitted Markov chain models.

In **Chapter 6**, we solve problems with analyzing target attributes that reside at a higher hierarchical level than the subgroup level (i.e., individuals are nested in measurement occasions). We answer **EMM-RQs B1, B2 and B3** by developing EMM for RCS data. We propose a generic approach for discovering subgroups displaying exceptional trend behavior by building quality measures on the concept of standard error. In addition, we give directions for how EMM-RCS can work with varying descriptor distributions, uneven spacing of measurements over time, fluctuating sample sizes and missing data. Our proposed methodology is evaluated on synthetic data and two public datasets. We furthermore propose a refinement operator for handling incomplete descriptors and perform two controlled experiments to assess its validity.

We demonstrate external validity of EMM-RCS in **Chapter 7** by discovering exceptional trends in adolescent alcohol use. We contribute to answering **D-RQ Public health** for var-

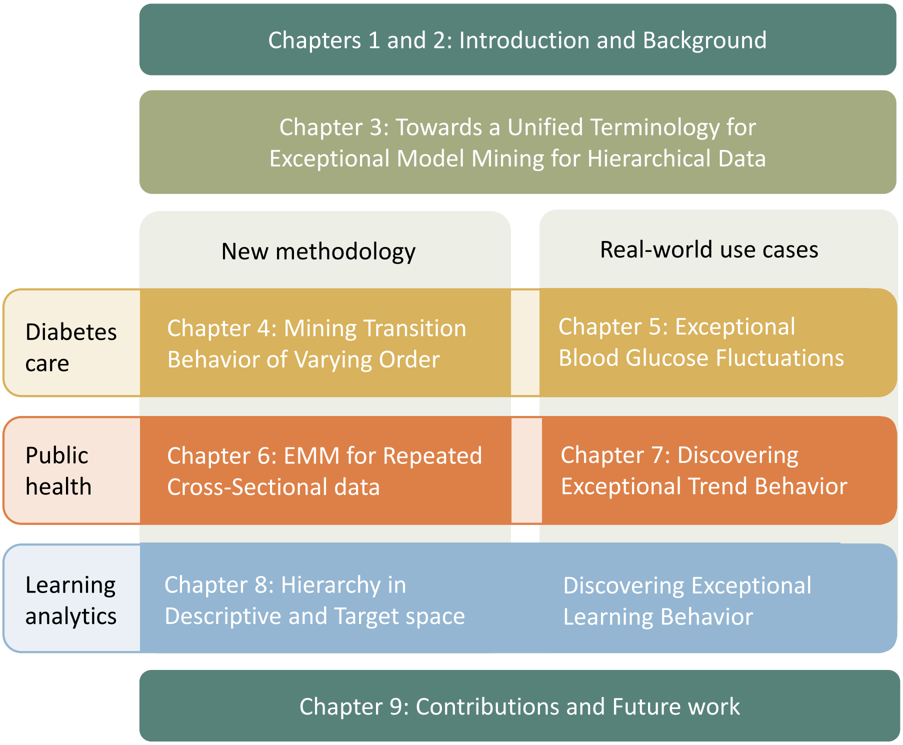

Chapters 1 and 2: Introduction and Background
Chapter 3: Towards a Unified Terminology for Exceptional Model Mining for Hierarchical Data
New methodology
Real-world use cases
Diabetes care
Chapter 4: Mining Transition Behavior of Varying Order
Chapter 5: Exceptional Blood Glucose Fluctuations
Public health
Chapter 6: EMM for Repeated Cross-Sectional data
Chapter 7: Discovering Exceptional Trend Behavior
Learning analytics
Chapter 8: Hierarchy in Descriptive and Target space
Discovering Exceptional Learning Behavior
Chapter 9: Contributions and Future work

Figure 1.2: Schematic outline of this dissertation.

ious types of trend deviations. In addition, we evaluate existing solutions for challenges such as reducing subgroup set redundancy and validation of the discovered subgroups. Together with domain experts, we demonstrate that EMM-RCS serves as a hypothesis-generating source that due to its exploratory nature works as a starting point in further understanding the interplay between socio-demographic variables and societal trends.

In **Chapter 8**, we analyze hierarchical data with nested observations in both descriptive and target space. Regarding challenges in descriptive space, we answer **EMM-RQ C1 and C2** and find that flattening data to long and wide flat-table data formats is incompatible with the hierarchical nature of data collected with digital assessment tools. As a solution, we propose the concept of aggregated descriptors and experimentally evaluate the approach of a Weighted Coverage Scheme (WCS) [117, 201].

Furthermore, we answer **EMM-RQ C3** by developing various quality measures based on the concept of least-squares. Our proposed method allows for the discovery of atypical subitizing patterns such as deviating initial reaction times, subitizing ranges, counting slopes or a combination of those. Indeed, domain experts confirm that our findings contribute to answering **D-RQ Learning analytics** since they support the belief that numerical processing competences strongly correlate with arithmetic skills.

Chapter 9 is the last chapter of this dissertation. There, we summarize the main conclusions of our work and discuss directions for future research.

1.5 Publications

Publications in scientific conference proceedings and journals that serve as core material for this dissertation:

- **Schouten, R. M., Engelen, B. L., Duivesteijn, W., and Pechenizkiy, M.** Towards a unified framework for Exceptional Model Mining for hierarchical data. *To be submitted to IJCAI 2025* (2024) [175].
  This manuscript serves as core material for **Chapter 3**.

- **Schouten, R. M., Bueno, M. L., Duivesteijn, W., and Pechenizkiy, M.** Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery 36* (2022), 379–413 [170].
  This publication serves as core material for **Chapters 4 and 5**.

- **Schouten, R. M., Duivesteijn, W., and Pechenizkiy, M.** Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In *Proc. SDM* (2022), pp. 585–593 [171] (supplementary material available through [172]).
  This publication and supplementary material serve as core material for **Chapter 6**.

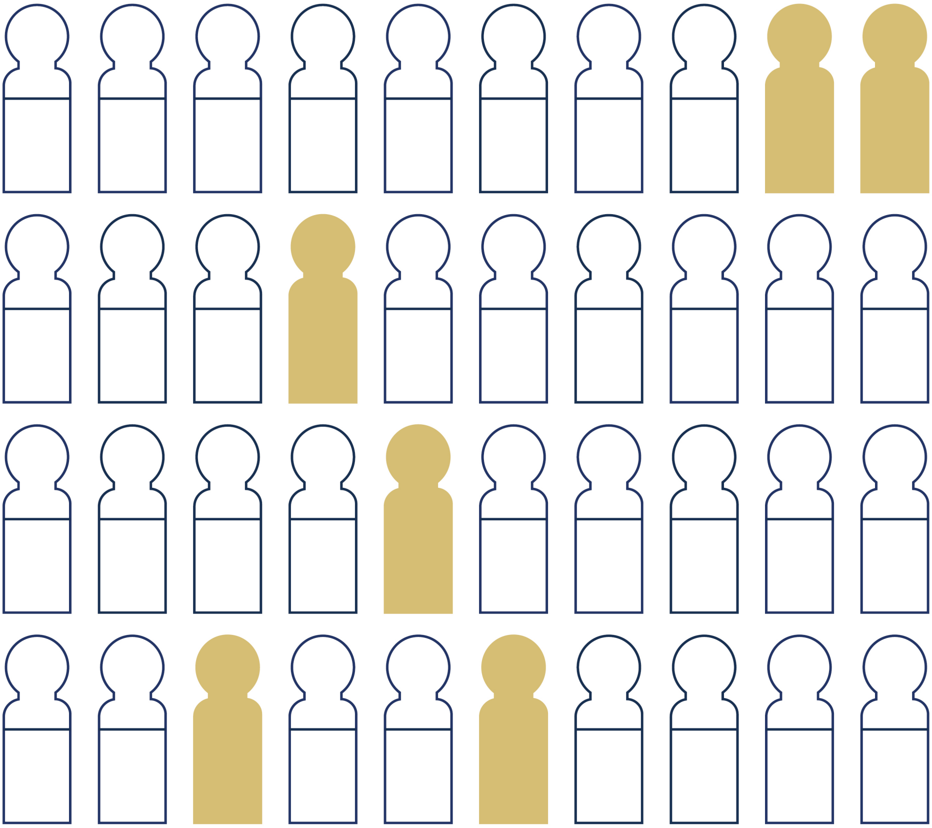
- **Schouten, R. M., Stevens, G. W., van Dorsselaer, S. A., Duinhof, E. L., Monshouwer, K., Pechenizkiy, M., and Duivesteijn, W.** Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining: Discovering exceptional trends in adolescent alcohol use in the Netherlands. *Accepted for presentation at BNAIC/BeNeLearn* (2024) [177].
  This manuscript serves as core material for **Chapter 7**.

- **Schouten, R. M., Duivesteijn, W., Räsänen, P., Paul, J. M., and Pechenizkiy, M.** Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression. In *Proc. ECML PKDD* (2024), p. 66–82 [174].
  This publication serves as core material for **Chapter 8**.

Publications to which I contributed that are not included in the dissertation:

- **Schouten, R. M.** On the role of prognostic factors and effect modifiers in structural causal models. *Accepted for poster presentation at Causal Representation Learning Workshop NeurIPS* (2024) [169].
- **van den Berg, N. T., Broekgaarden, B. O., Mahieu Dionysia, P., Martens, J. G., Niederle, J., Schouten, R. M., and Duivesteijn, W.** Generating MNAR missingness in image data, with additional evaluation of MisGAN. *Accepted for presentation at BNAIC* (2024) [197].
- **Schouten, R. M., Taşcău, V., Ziegler, G. G., Casano, D., Ardizzone, M., and Erotokritou, M. A.** Dropping incomplete records is (not so) straightforward. In *Proc. IDA* (2023), pp. 379–391 [178].
- **Verhaegh, R. F. A., Kiezebrink, J. J. E., Nusteling, F., Rio, A. W. A., Bendicsek, M. B., Duivesteijn, W., and Schouten, R. M.** A clustering-inspired quality measure for exceptional preferences mining – design choices and consequences. In *Proc. DS* (2022), pp. 429–444 [204].
- **van der Haar, J. F., Nagelkerken, S. C., Smit, I. G., van Straaten, K., Tack, J. A., Schouten, R. M., and Duivesteijn, W.** Efficient Subgroup Discovery through Auto-Encoding. In *Proc. IDA* (2022), pp. 327–340 [198].
- **Schouten, R. M., and Vink, G.** The dance of the mechanisms: How observed information influences the validity of missingness assumptions. *Sociological Methods & Research 50* (2021), 1243–1258 [179].
- **IJsselhof, R. J., Duchateau, S. D., Schouten, R. M., Slieker, M. G., Hazekamp, M. G., and Schoof, P. H.** Long-term follow-up of pericardium for the ventricular component in atrioventricular septal defect repair. *World Journal for Pediatric and Congenital Heart Surgery 11*, 6 (2020), 742–747 [90].
- **IJsselhof, R. J., Duchateau, S. D., Schouten, R. M., Freund, M. W., Heuser, J., Fejzic, Z., Haas, F., Schoof, P. H., and Slieker, M. G.** Follow-up after biventricular repair of the hypoplastic left heart complex. *European Journal of Cardio-Thoracic Surgery 57*, 4 (2019), 644–651 [89].
- **Schouten, R. M., Lugtig, P., and Vink, G.** Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation 88*, 15 (2018), 2909–2930 [176].
- **Kappen, I. F., Bittermann, G. K., Schouten, R. M., Bittermann, D., Etty, E., Koole, R., Kon, M., Mink van der Molen, A., and Breugem, C. C.** Long-term mid-facial growth of patients with a unilateral complete cleft of lip, alveolus and palate treated by two-stage palatoplasty: Cephalometric analysis. *Clinical Oral Investigations 21* (2017), 1801–1810 [96].
- **de Vries, C. P., Schouten, R. M., van der Kuur, J., Gottardi, L., and Akamatsu, H.** Microcalorimeter pulse analysis by means of principle component decomposition. In *Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray* (2016), vol. 99055V, pp. 1699–1708 [43].

2



Background

In this chapter, we provide background information on how Exceptional Model Mining originated as a Local Pattern Mining framework. We introduce notation and definitions for EMM, discuss the beam search algorithm and give an overview of existing methods for reducing subgroup set redundancy and validating discovered subgroups.

2.1 Local Pattern Mining

Knowledge Discovery in Databases (KDD) aims to extract novel, useful and interesting knowledge from databases [76, 110, 209]. KDD can be distinguished from Data Mining (DM), where KDD is the "overall process of discovering useful knowledge from data, while data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data" [58, p.82]. Here, the term *pattern* is used to refer to both patterns and *models* extracted from data [58].

A further division of data analysis techniques can be made by distinguishing *predictive induction*, the set of techniques that induce models from class labeled data, which are then used to predict the class value of previously unseen examples, from *descriptive induction*, which aims to discover comprehensible patterns, typically induced from unlabeled data [110]. It goes almost unnoticed that in this distinction, the term *model* is used with respect to predictive induction while the term *pattern* connects to descriptive induction. Indeed, [110] distinguish two research communities: the Machine Learning (ML) community focuses on developing techniques for predictive induction and the DM community is involved with descriptive induction.

In both communities, *rule discovery* is an important aspect of providing information in a human interpretable way. A rule is a statement of a regularity in the form of "if [premise], then [conclusion]" [194, p.207]. In the DM community, an association rule is "an expression $X \Rightarrow Y$ where X and Y are a set of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y" [1, p.307]. As such, Association Rule Mining (ARM) [2, 113] builds on Frequent Itemset Mining (FIM) [73]. In the ML community, the focus is on building *sets* of classification and prediction rules [32, 34] that together describe the population well. We could refer to this process as the discovery of *strong* rules: descriptions that apply to numerous objects (i.e., transactions) in a database with few counterexamples [194].

At the start of the 21$^{\text{st}}$ century, the concept of *global* models that explain most of the instances in the data was opposed with that of *local* models or patterns [74, 139]. The idea is that global models aim to find patterns that cover majorities in the population. The recent advances in deep learning and foundation models take place in this context. Alternatively, we may describe a population by a background model + a set of local patterns + random background noise [74]. Based on this definition, experts agreed that [139, preface]:

- local patterns cover small parts of the data space,
- local patterns deviate from the distribution of the population,
- local patterns show some internal structure.

Altogether, local patterns are interesting as targets of discovery since they deviate from the background model and could therefore reveal unknown information. Hence, Local Pattern Mining (LPM) [10, 139] originated as a subdomain of DM that aims to discover interesting subsets in the dataset. These subsets are not just collections of data points but they are patterns that are interpretable for domain experts and can be generalized to unseen data points; we call them *subgroups*.

Examples of LPM frameworks are Contrast Set Mining (CSM) [12] and Emerging Pattern Mining (EPM) [48]. Contrast sets are "conjunctions of attributes and values that have different levels of support in different groups" [12, p.214]. CSM is a special case of ARM where the consequent is restricted to a variable whose values denote group membership [206]. Emerging patterns are "itemsets whose support increases significantly from one data set to another" [48, p.43]. Emerging patterns can be seen as association rules with an itemset in the rule antecedent and a fixed consequent: $Itemset \rightarrow D_1$, for a given dataset D_1 being compared to another dataset D_2 [110].

Where the DM community shifted their focus from ARM towards CSM and EPM, the concept of *local* patterns motivated the ML community to move from building sets of classification and prediction rules to building individual rules for exploratory data analysis and interpretation. One of the tasks that emerged from this process is Subgroup Discovery (SD) [102, 110, 117, 209]: the task of identifying interesting subgroups according to some property of interest (more below in Section 2.2). A subgroup description can be seen as the condition of a rule: $SubgroupDescription \rightarrow Class$. Together, CSM, EPM and SD are considered *supervised descriptive rule discovery* tasks [110].

Another task that displays the evolution from global to local models in the ML community was introduced by [154], who presented an algorithm involving classification trees where two trees are grown in opposite directions so that they are matched at their leaves. The approach is called *redescription mining*: the task of finding subgroups having several descriptions or equivalence relationships of the form $E \Leftrightarrow F$ where E and F are set-theoretic expressions of binary attributes (or, items) [64, 154]. The work was extended from binary to real-valued data [63] and to relational data [61, 62].

It is clear that LPM frameworks emerged from different directions and with different purposes. The distinction between methods may be small and sometimes appears only after a detailed studying of terminology and definitions. Furthermore, proposed LPM frameworks may originate as mixtures of other methods. For instance, [191] propose redescription model mining. Others utilize LPM to build global models [88, 105, 117].

Remark that LPM differs from DM methods such as clustering [212], which is a global analysis approach that aims to divide all data points into k distinct groups, and outlier detection, which aims to detect single data points that deviate from the global model. Traditionally, neither methods focus on discovering a shared characteristic or interpretable description of the selected clusters and outliers. For instance, in Section 4.3, [164] take an approach similar to us by discovering unusual sequences based on the concept of log likelihood, but without providing additional descriptions of the detected outliers.

Some interesting work exist that combines techniques from LPM with clustering or outlier detection. For instance, [213] propose Cluster-Grouping (CG) as a subtask in SD, clustering and classification. Furthermore, [111] deliver an explanation of detected outliers in arbitrarily oriented subspaces of the original attribute space and [107] use SD as a postprocessing step to obtain descriptions after regular outlier detection.

In this dissertation, we zoom in on an LPM framework that is known as Exceptional Model Mining (EMM) [54, 121]. But first, we discuss Subgroup Discovery (SD) [102, 209].

2.2 Subgroup Discovery

The task of Subgroup Discovery (SD) was defined as follows: “In subgroup discovery, we assume we are given a so-called population of individuals (objects, customers, ...) and a property of those individuals we are interested in. The task of subgroup discovery is then to discover the subgroups of the population that are statistically most interesting, are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest” [210, p.84-85]. In the form of a rule, an induced subgroup description is defined as $Cond \rightarrow Target_{value}$, where $Target_{value}$ is a value for a variable of interest for the SD task (also referred to as $Class$). $Cond$ is a conjunction of attribute-value pairs [65, 102, 103, 117, 182].

Several interesting survey papers for SD exist [8, 76, 80]. They all reveal an abundance of quality measures and search algorithms. A quality measure $\varphi : 2^{\Sigma} \rightarrow \mathbb{R}$ maps every pattern (subgroup) in the search space to a real number that reflects the interestingness of a pattern (Σ is the set of all possible attribute-value pairs). Depending on whether the target variable is binary, nominal, ordinal or numeric, many quality functions exist, varying from coverage and support, to precision, sensitivity and specificity, to weighted relative accuracy, mean gain, z-score and many more [8, 80].

A clear overview of subgroup search algorithms is provided by [76], who distinguish three groups: beam search based algorithms, exhaustive search algorithms and genetic algorithm based approaches. An anytime algorithm with guarantees was proposed by [15]. Interestingly, already since the start of the development of SD algorithms, special attention was given to SD in other than single-relation databases. Multi-relational SD was proposed by first defining “a designated object relation intended to be the master relation of the population of interest” [209, p.80]. Subsequently, sampling techniques filter the appropriate target data without joining all possible relations. Alternatively, [103] traverse the search space by using a set of aggregation functions such as count, average, max, and min, while additionally, a predefined relation graph that determines which joins are selected. Another valuable approach to performing multi-relational SD is by way of propositionalization: transforming a relational database into a single-data-table representation [211].

The domain of SD furthermore focused on exploiting background knowledge in the form of ontologies, called Semantic SD or Semantic DM [118]. An ontology is a hierarchy of concepts, such as the nesting of cities in regions in countries [202], or the nesting of protein binding sites in genes [118, 119, 140, 185]. The hierarchical structure is generally given by domain experts. Since every entity in the dataset has an associated value for each of the concepts, ontologies are valuable for traversing the search lattice efficiently.

2.3 Exceptional Model Mining

Another LPM stream that builds on the SD task is the framework of Exceptional Model Mining (EMM) [54, 121]. EMM is considered a generalization of SD where the interestingness of subgroups is determined over *target models* rather than a single target attribute. It aims to discover subgroups in the dataset where a model fitted to the subgroup is substantially different from a model fitted to the entire dataset. EMM commonly takes into account ≥ 2 target attributes whereas SD uses 1.

The idea of EMM was introduced by [121] and further developed by many authors. In 2016, a cohesive overview of EMM terminology, search strategies and instances was provided by [54]. Examples of target models include correlation [121], association [54], linear regression [53] and Bayesian Networks (BNs) [56].

In this chapter, we provide a short introduction to EMM. We particularly focus on notation and terminology (Section 2.3.1) and beam search as the search algorithm of our choice (Section 2.3.2). We give an overview of existing techniques for reducing subgroup set redundancy and for validating the discovered subgroups in Section 2.3.3. We discuss specific methods that are relevant and related to our proposed methodologies in the respective chapters of this dissertation. Furthermore, **Chapter 3** extensively discusses related work on EMM for hierarchical data. Here, we present traditional EMM definitions.

2.3.1 Traditional notation and terminology

Traditionally, EMM assumes a dataset Ω to be a bag of n records $r \in \Omega$ of the form

$$r = \left(a_1, \ldots, a_k, \ell_1, \ldots, \ell_m\right), \tag{2.1}$$

where k and m are positive integers [54].[1] In EMM, we call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. For SD, $m = 1$, whereas for EMM, typically $m \geq 2$. Figure 2.1 gives a schematic illustration of a dataset for $n = 5$, $k = 4$, and $m = 3$. Whether an attribute is deployed as descriptive or target attribute is generally based on the application domain.

The descriptive attributes are used to describe and discover subgroups of records. A subgroup is defined using descriptions; a description is a Boolean function $D : \mathscr{A} \to \{0, 1\}$ which covers a record r^i if and only if $D\left(a_1^i, \ldots, a_k^i\right) = 1$. Here, $\mathscr{A}$ is the collective domain from which the full set of descriptors is taken; a Cartesian product of the domains of each individual descriptor. Consequently, a subgroup is defined as follows:

Definition 2.1 (Subgroup cf. [54])**.** *A subgroup corresponding to description D is the bag of records $G_D \subseteq \Omega$ that D covers:*

$$G_D = \left\{r^i \in \Omega \mid D\left(a_1^i, \ldots, a_k^i\right) = 1\right\}.$$

The complement contains all records that are not covered; $G^C = \Omega \setminus G_D$.

In EMM, the choice of *description language* $\mathscr{D}$ is free, though generally we let the description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathscr{A}_j$ of the respective attribute a_j. For instance, for discrete variables the selector may be an attribute-value pair ($a_j = v$); for continuous variables it could be a range of values ($w_1 \leq a_j \leq w_2$) [54, 135, 171].

[1] We follow up on notation defined by [54], but there are no major differences with other prominent work on SD and EMM, such as [8, 102, 121].

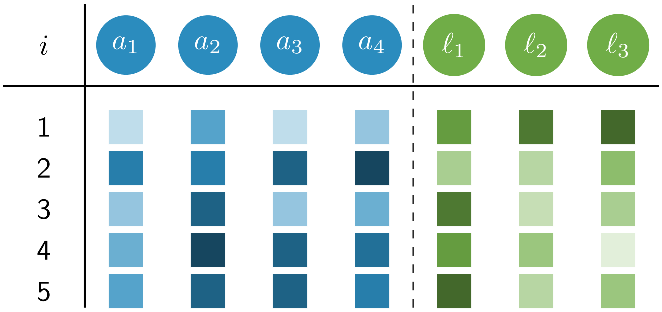


Figure 2.1: Schematic illustration of dataset Ω as commonly used in EMM.

We aim to discover the descriptions for which the subgroups display exceptional behavior on a target model, fitted to a set of target attributes. The target model of interest and type of target variables generally depend on the application. A straightforward example is a correlation model, where we are interested in the linear association between two numeric targets ℓ_1 and ℓ_2 as measured by the correlation coefficient ρ. A possible aim is to discover subgroups with exceptionally deviating correlation coefficients [121].

Formally, we quantify exceptionality using a quality or interestingness measure. A quality measure quantifies the difference between behavior in the subgroup and some reference behavior, usually the subgroup's complement:

Definition 2.2 (Quality Measure cf. [54])**.** *A quality measure is a function $\varphi : \mathscr{D} \rightarrow \mathbb{R}$ that assigns a numerical value to a description D.*

For correlation as target model, a possible quality measure is $\varphi_{scd} = 1 - p$ where p is the p-value of a statistical hypothesis test that evaluates $H_0 : \rho^G = \rho^{G^C}$ against $H_1 : \rho^G \neq \rho^{G^C}$ [121]. The larger the difference between the correlation coefficients of the subgroup and the subgroup's complement, the smaller the p-value and the higher the quality value φ_{scd}. Subgroups with higher quality values are considered more exceptional than subgroups with lower quality values. The task of EMM is to discover these subgroups.

Quantification of exceptionality using quality measures is usually not sufficient to discover interesting subgroups. After all, the most extreme deviations from the norm can be found in very small subsets of the data; these are generally not the most interesting for domain experts. Therefore, we somehow want to incorporate the subgroup size when evaluating candidate subgroups, for instance by multiplying φ with n or $\sqrt{n}$. Another interesting approach is to multiply with the entropy function [54, cf.] The entropy φ_{ef} is maximal (close to 1) when subgroup and complement both contain 50% of the data rows; it is smallest (close to 0) when subgroup or complement are (almost) empty.

Remark that statistical tests generally require independence of observations. Although formal EMM definitions are agnostic about the origin of the data and whether or not record r^i is independent from record r^j, most commonly, dataset Ω contains n independent observations, drawn from multivariate state-space $\mathscr{A} \times \mathscr{L}$. It is then possible to format data as in Figure 2.1, to use statistical tests to quantify the exceptionality of subgroups and to directly compare G to G^C.

For the same reason, in traditional SD and EMM, entities may be described as *records, cases, individuals, transactions, objects* or *observations* (at this point in this dissertation, we have encountered all of these already). Since every independent draw is another entity of the same entity type, the terms used to describe that type are interchangeable. However, in the context of hierarchical data, the data contains observations on multiple entity types and the relative positioning of these entity types is important for whether or not observations are independent. In this context, it is challenging to directly compare G to G^C and to use statistical tests to quantify exceptionality. Alternative methods are needed.

2.3.2 The beam search algorithm

The task of EMM is to effectively search through the space of candidate subgroups to find the top-q best-scoring subgroups [54]. Conform [54], the formal task definition of EMM can be defined as:

Definition 2.3 (Problem statement top-q Exceptional Model Mining)**.** *Given a dataset Ω, description language $\mathscr{D}$, quality measure φ, positive integer q, and a set of constraints $\mathscr{C}$, the top-q task is to find the list $\{D_1, ..., D_q\}$ of descriptions in the language $\mathscr{D}$ such that*

- $\forall 1 \leq i \leq q$: D_i *satisfies all constraints in* $\mathscr{C}$*;*
- $\forall i, j : i < j \Rightarrow \varphi(D_i) \geq \varphi(D_j)$*;*
- $\forall D \in \mathscr{D} \setminus \{D_1, ..., D_q\}$: D *satisfies all constraints in* $\mathscr{C} \Rightarrow \varphi(D) \leq \varphi(D_q)$.

Possible constraints include a minimum subgroup size, a constraint on the quality value or the application of pruning and anti-redundancy techniques (see Section 2.3.3).

Many search algorithms exist; some of them developed for particular kinds of exceptional behavior [17, 24, 123, 131], others for particular data types [118, 132, 138, 158]. Nevertheless, most EMM methodologies consider the search space to be a general-to-specific search lattice where relations are formed by subgroups whose descriptions differ by conjoining of a single additional selection condition (i.e., more general (specific) subgroup descriptions occur higher (lower) in the lattice and cover more (fewer) instances). The core difference between most search algorithms is the way in which they traverse the search lattice; given a (candidate) subgroup description, the selection of subgroup members is comparable for many search algorithms.

In all chapters in this dissertation, we choose *beam search* [54, Algorithm 1] as our search algorithm of choice. We have three important reasons. First, beam search discovers exceptionally behaving subgroups using a heuristic search strategy. It is an intuitive method that can easily be understood at a conceptual, non-technical level. This makes beam search perfect when working together with experts from various domains. Furthermore, given a deterministic quality measure, beam search is deterministic; with the same data and parameter settings, the algorithm returns the same top-q subgroups. Third, beam search is

Algorithm 1 Beam Search Algorithm cf. [54, Algorithm 1]

```
    Input Dataset $\Omega$, quality measure $\varphi$, refinement operator $\eta$, beam width $w$,
      beam depth $d$, result set size $q$, constraints $\mathscr{C}$
    Output PriorityQueue resultSet
 1: candidateQueue ← new Queue;
 2: candidateQueue.enqueue({});
 3: resultSet ← new PriorityQueue($q$);
 4: for (Integer level ← 1; level ≤ $d$; level++) do
 5:     beam ← new PriorityQueue($w$);
 6:     while (candidateQueue $\neq \emptyset$) do
 7:         seed ← candidateQueue.dequeue();
 8:         set ← $\eta$(seed);
 9:         for all (desc ∈ set) do
10:             quality ← $\varphi$ (desc);
11:             if (desc.SATISFIESALL($\mathscr{C}$)) then
12:                 resultSet.insert_with_priority(desc,quality);
13:                 beam.insert_with_priority(desc,quality);
14:     while (beam $\neq \emptyset$) do
15:         candidateQueue.enqueue(beam.get_front_element());
16: return resultSet;
```

not restricted to particular descriptors, target models or quality measures; it can be applied as long as there exists a set of descriptors $a_1, \ldots, a_k$ which can be of any type (binary, nominal, ordinal and/or numerical) and it can work with any quality measure defined over target attributes.

The beam search algorithm is shown in Algorithm 1 (cf. [54]). In essence, beam search performs a level-wise search of d levels (line 4). At each level, w promising descriptions are selected into the beam (line 13); these descriptions are taken to the next level (line 15) and refined further (line 8). Refinement of candidate subgroups occurs by conjoining single additional selection conditions to existing subgroup descriptions (line 8). Following [54], the time complexity of the beam search algorithm is:

$$\mathscr{O}(dwkn(c + M(n,m) + \log(wq))). \tag{2.2}$$

Here, $M(n,m)$ is the time complexity of evaluating the quality of a target model on n records and m targets; c is the cost of comparing two models. After a pre-specified number of levels d, the top-q subgroups are returned.

2.3.3 Pruning and validation techniques

Considering that our methodologies aim to answer RQs that impact people in the real world, and considering that practitioners and policy makers may build upon these results, it is crucial that we validate the discovered subgroups. Therefore, throughout this dissertation, we apply the following two techniques: 1) we construct a Distribution of

False Discoveries (DFD) as proposed by [52] and 2) we apply a minimum improvement (MI) threshold [13]. Both approaches are discussed in detail below.

In addition, while discovering exceptional subgroups, we distinguish the process of local pattern discovery from the process of discovering a *set* of local patterns [104]. The former ensures that each $D \in \{D_1, ..., D_q\}$ is interesting on its own, while the latter takes into account overlap and similarities between selected subgroups. Subgroup set discovery aims to deliver a subgroup set $\{D_1, ..., D_q\}$ that is somehow optimal *as a set*. Generally, the two processes are executed sequentially. First, we discover interesting subgroups. Second, pruning and post-processing techniques are used to remove redundant or non-significant subgroups [104, 105, 205].

An alternative approach is to incorporate pruning techniques while creating the beam, at each level in the search. This approach is computationally more efficient and increases the diversity of the final pattern set [200, 201]. Specifically, we may differentiate between redundancy of subgroup descriptions (subgroups having similar descriptions), redundancy of subgroup covers (subgroups covering the same set of individuals) and redundancy of exceptional models (subgroups having the same exceptionality score) [201].

In this dissertation, we apply three anti-redundancy techniques: Description-Based Selection (DBS) [201] to reduce redundancy of subgroup descriptions and exceptional models, a Weighted Coverage Scheme (WCS) [117, 201] to increase the variation in subgroup cover, and Dominance-Based Pruning (DBP) [201] to remove conditions that decrease the quality of the subgroup. Table 2.1 provides an overview of the anti-redundancy and validation techniques that are used in the chapters of this dissertation. In Sections 7.5 and 8.7.1, we perform additional experiments to evaluate the effects of these techniques. We now discuss DBS, WCS, DBP, DFD and MI in detail.

Description-based selection

With Description-Based Selection (DBS) [201], we evaluate the descriptions and quality values of candidate subgroups before determining the beam. The strategy greedily selects subgroups by comparing each candidate to the subgroups already selected (starting with the subgroup with highest quality). Specifically, subgroups are skipped if they have 1) equal quality and 2) the same selection conditions except for 1 condition. We aim to keep the candidate subgroup with the most general description. For instance, these are descriptions with fewer conditions or descriptions based on numerical attributes where the numerical range is wider.

We distinguish fixed-size DBS, where candidate subgroups are evaluated until a fixed beam size is reached, from variable-size DBS, where each description attribute is allowed to occur a fixed number of times in a condition in a subgroup set [201]. In this dissertation, we always apply fixed-size description-based selection. Most commonly, we set the beam size to $2w$ (where w is the original beam size without performing DBS). The fixed beam size should not be too large, to reduce the computational cost of subsequent procedures (WCS and DBP, see below); the beam size should not be too small, to warrant that a sufficient number of candidate subgroups is available.

Remark that rounding of quality values influences whether or not the first requirement in DBS is met (subgroups are skipped if they have equal quality). Without sufficient rounding of quality values, is it likely that none of the subgroups will be removed. At the same time, if the rounding procedure is too strict, we may not be able to distinguish between candidate subgroups.

Weighted coverage scheme

A Weighted Coverage Scheme (WCS) was initially introduced by [117] when they incorporated the cover of candidate subgroups into the Weighted Relative Accuracy (WRAcc) measure. The idea was generalized by [201], who proposed to weigh the quality value of each candidate subgroup based on the number of times individuals are already covered by other subgroups. Like DBS, a WCS is used at every search level and aims to create a diverse beam.

A WCS selects the desired number of subgroups in iterations. Initial counts for all individuals are 0. Consequently, in the first iteration, all individuals are equally likely to be covered (equal counts of 0, weights are all 1) and the subgroup with the highest quality value is selected. Second, the counts of the individuals covered by the first subgroup are increased with 1. This changes the weights of the candidate subgroups and reduces their quality value depending on how many individuals are covered. After re-calculating all quality values, the current best subgroup is selected, counts are updated, weights are updated and the process iterates until the desired w (in case of selection of the beam) or q (in case of preparing the final result list) subgroups are selected.

Table 2.1: An overview of the anti-redundancy and validation techniques used in this dissertation, given per chapter and per experiment. Parameter settings are provided in the respective chapters.

Chapter	Experiment	DBS	WCS	DBP	DFD	MI	Code
4	Synthetic transition						A
	Synthetic starting	✓	✓	✓			A
	Synthetic sensitivity						A
	MovieLens	✓	✓	✓			A
5	DIALECT-2	✓	✓	✓			A
6	Synthetic	✓	✓	✓			B
	MD Brexit	✓	✓	✓			B
	MD HBSC/DNSSU	✓	✓	✓	✓		C
	Eurobarometer	✓	✓	✓	✓		B
	Brexit	✓	✓	✓	✓		B
7	HBSC/DNSSU	✓	✓	✓	✓	✓	C
8	Exploration WCS		✓				D
	FUNA		✓				D

[A] https://github.com/RianneSchouten/simulations_markov_chains_emm/

[B] https://github.com/RianneSchouten/EMM_RCS/

[C] https://github.com/RianneSchouten/AlcoholTrends_HBSCDNSSSU_EMM/

[D] https://github.com/RianneSchouten/FUNA_EMM/

The WCS proposed by [201] updated the quality values by means of multiplicative weighted coverage. Then, in each iteration, the weight of candidate subgroup G is:

$$\tau^G_{\text{mult}} = \frac{1}{n^G} \sum_{i=1}^{n^G} \gamma^{c_i}, \tag{2.3}$$

where c_i counts the number of times that individual i is already covered by previously selected subgroups (the higher the count, the lower τ^G_{mult} and the more the reduction in quality). Parameter value $\gamma \in [0,1]$ reflects how strict our weighting regime should be: at one end of the axis we remove covered instances from the dataset before choosing the next pattern ($\gamma = 0$); at the other end we give equal weights to all individuals independent of the number of times they have been covered ($\gamma = 1$). The smaller the value of γ, the more strict the weighting regime.

As an alternative, additive weighted coverage was proposed [117]. Here, each subgroup is weighted with:

$$\tau^G_{\text{add}} = \frac{1}{n^G} \sum_{i=1}^{n^G} \frac{1}{c_i + 1}, \tag{2.4}$$

In this dissertation, we apply fixed-size multiplicative WCS with $\gamma = 0.9$. In Section 8.7.1, we explore the interaction effect between γ and search depth d. Recall that in this dissertation, we apply DBS and WCS sequentially. This means that WCS is only applied to the set of candidate subgroups that have survived the DBS.

Dominance-based pruning

Beam search can lead to subgroup descriptions where a certain subset of conditions has a higher quality value than the full description itself [201]. This may happen especially with binary variables. For instance, the description *sex = female* ∧ *ethnic group = western* may not appear in the results list because the individual descriptions *sex = female* and *ethnic group = western* did not have high quality (and were therefore not included in the top w subgroups at level 1 of the search). However, the description *age ≤ 14* ∧ *sex = female* ∧ *ethnic group = western* may have been discovered (because *age ≤ 14* was selected into the beam at $d = 1$), but with lower quality than the subgroup description based on the two conditions on descriptors *sex* and *ethnic group*. Here, we say that a subset of conditions *dominates* the original description [201].

We apply a form of Dominance-Based Pruning (DBP) where we evaluate the quality of all subgroups that can be formed based on the subsets of the conditions of the descriptions of the top-q subgroups. We thus apply DBP after the result list has been determined. In the situation that a certain subset of conditions has a higher quality value, we will adopt it as a new description and place it in the results list. Obviously, that means that the subgroup at position q will be removed from the list. We consider DBP to be a form of anti-redundancy as well as subgroup validation because it removes those conditions from descriptions that do not substantially add to the quality of the subgroup.

Distribution of False Discoveries

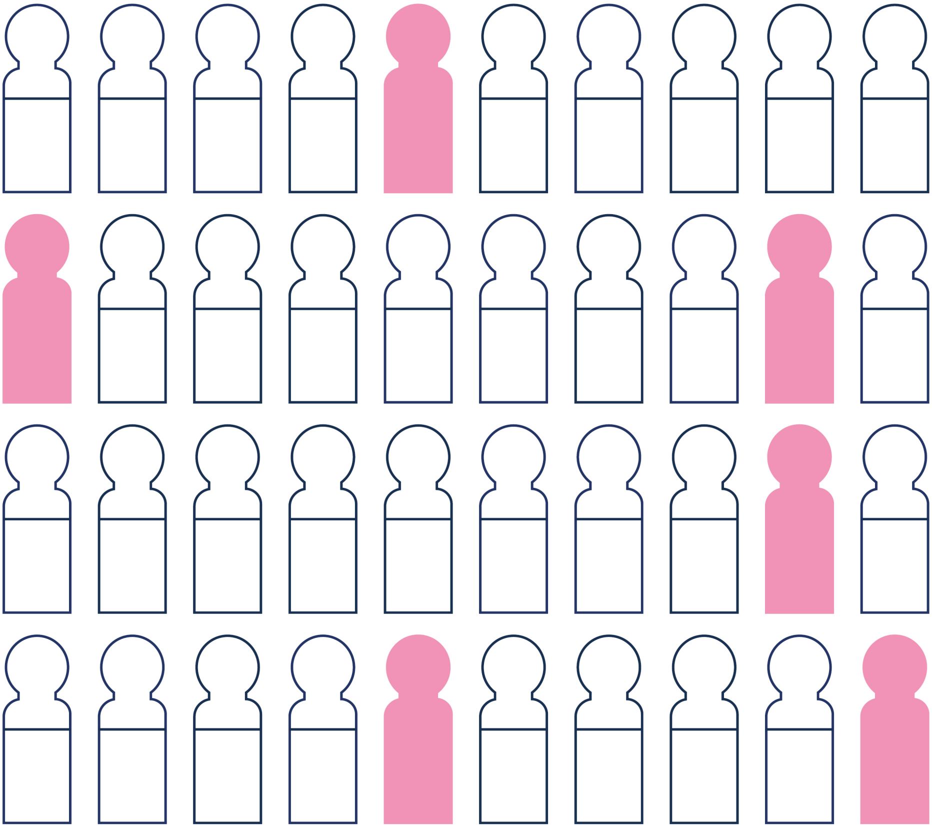
To validate the significance of the discovered subgroups, the quality values of all top-q subgroups are compared against a null *Distribution of False Discoveries* (DFD) [cf. 55]. The DFD is constructed as follows. First, we randomly swap the values of descriptors between individuals while keeping the information on targets intact. For instance, for our study of adolescent alcohol use in Section 7.5, we exchange the values of adolescent r^i (who has age 13, life satisfaction 8 and a Dutch ethnicity) with the values of adolescent r^j (who has different values on these variables). Here, if adolescent r^j was drinking alcohol (the target attribute), we keep that information intact, which makes that drinking alcohol is now associated with an age of 13, a life satisfaction of 8 and a Dutch ethnicity. In other words, the swap randomization removes the correlation between descriptors (socio-demographic variables) on the one side and targets (alcohol use) on the other side, while keeping the distributions intact.

Next, we perform beam search on the swapped randomized dataset and store the quality value of the best-scoring subgroup. This subgroup and associated quality value should be considered a false discovery because the relation between the socio-demographic variables and alcohol use does not truly exist in the swap-randomized dataset. By repeating the procedure m times, we obtain m quality values which are used to construct a null distribution; the DFD. Under the assumption that m is sufficiently large, the mean of the quality values follows a normal distribution. We then run beam search on the non-swapped, original dataset and compare the quality values of the top-q subgroups against the DFD by means of a Z-test. The chosen values for m and the significance level α_{DFD} are different in different chapters of this dissertation. If m is not sufficiently large, a p-value can be calculated non-parametrically by means of ranks.

Minimum improvement threshold per condition

Lastly, we ensure that all conditions of a subgroup description substantially add to the subgroup's quality by adding a Minimum Improvement (MI) threshold [13]. While DBP removes conditions that *de*crease the quality value of a subgroup, here we aim to evaluate whether every condition results in a substantial *in*crease in quality. Although MI does not fully dismiss the probability that any discovered pattern is a chance artifact [205], we consider the method valuable for assessing whether an increase in quality is practically meaningful and relevant.

3



Towards A Unified Terminology

Subgroup Discovery (SD) and Exceptional Model Mining (EMM) aim to discover subgroups in a population that somehow behave exceptionally. Both SD and EMM are continuously evolving research areas, especially towards the discovery of subgroups in datasets that have some kind of hierarchical structure that deviates from the typical row-by-column table format. Although formal EMM definitions do not assume data points to be independent, we observe a problem with the way current terminology is deployed for non-IID data. On the one hand, overlooking a possibly nested data structure increases the risk of unintentionally finding nested subgroups. On the other hand, understanding similarities and differences between existing work is challenging because data structures and proposed methods are described using a variety of terms and definitions; this disturbs the process of mapping application questions to available EMM methodologies and may lead to reinventing the wheel. As a solution, we provide a unified terminology for EMM with hierarchical data, we propose the notion of subgroup level and we classify existing literature based on whether descriptive or target attributes reside at a lower, the same or a higher hierarchical level than the subgroup level.

The contents of this chapter will be submitted to IJCAI 2025.

3.1 Introduction

Subgroup Discovery (SD) [80, 102] and Exceptional Model Mining (EMM) [54, 121] are local pattern mining frameworks, seeking interpretable subgroups in a population that somehow behave exceptionally (see Section 2.3). Both SD and EMM are continuously evolving research areas in the domain of Data Mining (DM) [8, 76, 110, 174]. In the rest of this chapter, we will use the term EMM as overarching term for both SD and EMM, unless a clear distinction has to be made.

One of the directions in which EMM evolves, is the discovery of subgroups in datasets that have some kind of hierarchical structure; the structure deviates from the typical row-by-column table format. For instance, we have Exceptional Subgraph Mining [19, 98], Sequential EMM [131, 138], and Temporal EMM [26]. Indeed, in many applications, data points are *not* Independent and Identically Distributed (IID) and it is valuable to develop data mining methods that do not assume IID data.

However, in the EMM literature, we observe a problem with the way traditional terminology is applied to non-IID data. Although formal definitions in EMM do not assume data points to be independent, in practice much of the work has focused on single-table row-by-column data (more in Section 3.5). As a consequence, we see a potential risk of confusion when the EMM framework further develops towards non-IID data. For instance, we come across work that seems to assume IID data but experiments on data with a nested (dependent) structure [10, 50], and we find work that seems to apply the same target model but in reality discovers subgroups of entities that reside at different hierarchical levels [124, 170].

Furthermore, we notice that existing work on EMM with hierarchical data use a variety of terms and definitions that makes it hard to compare their approaches. For instance, Sequential EMM [131, 138] is proposed for discovering subgroups of subsequences of (binary) itemsets and Temporal EMM [26] deploys Dynamic Bayesian Networks (DBNs). As another example, [147, 158] both work with hierarchical attributes, but for [158] these exist in target space, whereas [147] operates in descriptive space. Moreover, the hierarchical discretization technique proposed by [147] relates to Semantic SD [118, 119, 140], who use knowledge graphs and ontologies to traverse the search space of candidate subgroups. These kind of differences and similarities between existing work are not immediately obvious based on a quick reading and prevent further development of EMM towards important research directions.

As a solution to this problem, we provide a unified terminology of existing EMM literature for hierarchical data. We first define hierarchical data as a collection of measurements taken from different types of entities, where the measurements of one type of entity are nested in the measurements of another type of entity. We then define the hierarchical level of the type of entity for which the subgroup should be formed as the *subgroup level* and distinguish between descriptive and target attributes that abide at a lower, the same or a higher hierarchical level. Based on this, we categorize existing literature into 9 classes: all combinations of $\{\text{lower}, \text{same}, \text{higher}\} \times \{\text{lower}, \text{same}, \text{higher}\}$. Naturally, EMM with IID data will fall in the same-same category.

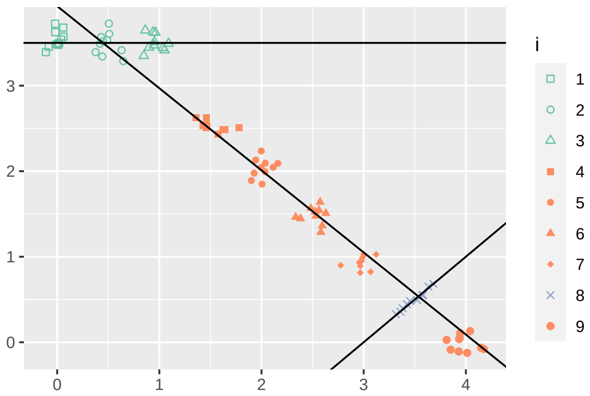


Figure 3.1: Toy example: discovering subgroups with exceptional linear relations in hierarchical data.

To illustrate the importance of taking into account a possibly nested data structure, we give a toy example in Figure 3.1. Regression is frequently used as target model in EMM [53, 54, 121, 143]: the goal becomes to find concisely defined subgroups where two (or more) target columns of the dataset display an unusual linear relation, as measured by (linear) regression. The figure displays a scatter plot of the target space; subspaces of the dataset contributing to subgroup definition are not shown. The data contains $j = \{1, 2, \ldots, 10\}$ measurements per person $i \in \{1, 2, \ldots, 9\}$. The marginal means of x and y differ per person, but overall the linear regression model has a negative coefficient (whether we average over $N = 90$ data points or $n = 9$ persons). Now, if our interest is in finding subgroups of persons with an exceptional relation between x and y, an EMM procedure ideally selects persons $i \in \{1, 2, 3\}$; they have a horizontal regression equation. However, if we construct the data as a 90 × 2 flat table, the search algorithm will rather select all the data points belonging to person $i = 8$ and consider that to be the top-1 exceptional subgroup. Then, it is easy to conclude that there exists a subgroup with a strongly positive relation, but really such a relation only exists *within one person*, and not within a subgroup of persons. Hence, a more careful consideration of the relative hierarchical level on which our observations, our subgroup definitions, and our target models lie, has the potential to lead us to more interesting conclusions.

In sum, our contributions include: 1) a unified terminology for EMM for hierarchical data, 2) a classification of existing EMM literature using this terminology, including a category for IID data and 3) an overview of interesting future research directions.

3.2 Background

Generally in DM, we start with the availability or definition of a dataset. In this section, we take one step back and look at how a dataset is a collection of observed measurements and a degree of uncertainty around those measurements. We then give a short recap of traditional terminology in EMM (more can be found in Section 2.3) and explain why existing definitions intuitively suggest that data is IID rather than non-IID.

3.2.1 Independent and identically distributed sequences

Given a probability space $(\Omega, \Sigma, \mathbb{P})$, A Random Variable (RV) $X : \Omega \to \mathscr{X}$ is a function that maps the sample space to a state space. Once an RV is introduced, the sample space Ω is no longer important; it suffices to list the possible values of X and their corresponding probabilities by means of distribution functions. Probability distribution $p(X)$ is a probability mass function (pmf) or probability density function (pdf), depending on whether the state space is discrete or continuous. Naturally, many distinct RVs may be defined over the same sample space. We call these *multivariate RVs* and denote them as a vector $\mathbf{X} = (X_1, \dots, X_d)$ with state space $\mathscr{X} = \mathscr{X}_1 \times \dots \times \mathscr{X}_d$ [45].

A sequence of RVs $X^1, \dots, X^N$ is *Independent and Identically Distributed*, also written as i.i.d, iid, or IID, when every X^i has the same distribution and X^i is independent of X^j, that is, X^i does not influence the value of X^j and vice versa.[1] The advantage of an IID sequence is that we can discover characteristics of the distribution of X^i, such as the mean $\mathbb{E}[X^i]$ or variance $\mathbb{V}[X^i]$, by averaging over the N RVs [45].[2]

When working with a dataset D, we commonly assume $D = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ to be an IID sample of size n. We may also write that we consider $\mathbf{X}$ to be a vector of RVs taking values in $\mathscr{X}$ and that dataset $D = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^n\}$ is a collection of n independent draws from $\mathscr{X}$. In both formulations, RV X^i is "the result of the i^{th} repetition of a particular measurement or experiment" [45, p.181]. When defining data, we consider such measurement or experiment to be one observations of a univariate or multivariate RV. However, the central limit theorem also allows us to extend the principle to summary statistics or to the parameter estimates of a descriptive (or predictive) model.[3]

3.2.2 Local patterns

Summary statistics or descriptive models may not adequately convey all interesting information on a target variable Y. Two famous illustrative examples are Anscombe's quartet [5] and Simpson's paradox [183]. Both examples demonstrate that efficiently estimating a population parameter θ not necessarily ensures that estimate $\hat{\theta}$ provides a good description of that population.

[1] Note the use of a superscript to clearly indicate a difference with the notation used for a multivariate RV. In other words, X^i may be a univariate or multivariate RV.

[2] The law of large numbers and the central limit theorem state that as $N \to \infty$, the difference between the estimated values and the true values approximates 0; subtle differences exist between the law of large numbers and the central limit theorem.

[3] From a sampling theory perspective, we aim to summarize the behavior of a target population $U = \{1, 2, \dots, N\}$ of N elements by means of population parameters; these are numerical indicators that depend on the values $Y_1, Y_2, \dots, Y_N$ of a target variable. Formally, a sample is a sequence of indicators $a = (a_1, a_2, \dots, a_N)$ where a_k is the number of times that element k is selected in the sample. A sampling design p assigns to every possible sample a a probability $p(a)$ of being selected. We then summarize the behavior of the target population using estimators; an estimator t for a population parameter θ is called an unbiased estimator if $\mathbb{E}[t] = \sum_{a \in A} t(a)p(a) = \theta$. The variance of an estimator t is equal to $\mathbb{V}[t] = \mathbb{E}[(t - \mathbb{E}[t])^2]$. Hence, by definition, an estimator t is an RV itself. Consequently, an estimator has a normal distribution with as expected value the population parameter θ and $\mathbb{V}(t)$ as the variance. This approximation works better as we draw more samples $a \in A$; Of course, in practice, we only draw one sample of size n and estimate the variance of estimator t using the sampled values. The variance can be estimated more precisely when the sample size n increases [22].

In EMM, we distinguish global patterns, that cover most instances in the data, from local patterns, that cover small parts of the data space, deviate from the distribution of the population and show some internal structure [54, 74, 139]. The framework of EMM aims to 1) discover subsets in the population somehow behave exceptionally and 2) describe these subsets in interpretable terms; they are then called *subgroups* [54].

Formal terminology and definitions for EMM were given in Section 2.3. Essentially, EMM assumes a dataset Ω to be a bag of n records $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m), \tag{3.1}$$

where k and m are positive integers. In EMM, we call $a_1, \ldots, a_k$ the *descriptive attributes* or *descriptors* of r, and $\ell_1, \ldots, \ell_m$ the *target attributes* or *targets* of r. For SD, $m = 1$, whereas for EMM, typically $m \geq 2$. Definitions of a subgroup (Definition 2.1) and a quality measure (Definition 2.2) were given in Section 2.3.

A schematic illustration of a dataset commonly used in EMM was given in Figure 2.1 for $n = 5$, $k = 4$ and $m = 3$. It is important to note that formal EMM definitions are agnostic about the origin of the data: whether or not record r^i is independent from record r^j, and whether or not the data should be formatted as a flat table. However, most commonly, dataset Ω is formatted as in Figure 2.1: the table rows contain records "$i \in D$, referred to as individuals, taken from a domain D" [121, p.3] and columns represent attributes. It then makes sense to consider each record to be an independent observation $\mathbf{x}^i = (\mathbf{a}^i, \boldsymbol{\ell}^i) = (a^i_1, a^i_2, \ldots, a^i_k, \ell^i_1, \ell^i_2, \ldots, \ell^i_m)$, drawn from a multivariate state space $\mathscr{X} = \mathscr{A} \times \mathscr{L}$ where "descriptors are taken from an unrestricted domain $\mathscr{A}$" [54, p.52] and targets from $\mathscr{L}$. In Section 3.5.1, we give three other arguments for our premise that most existing EMM methodologies assume IID data.

3.2.3 Hierarchical data

In this dissertation, we consider hierarchical data, a form of non-IID data where measurements of one set of RVs are nested in the measurements of another set of RVs. The concept is well-known in the social and biomedical sciences and stems from the idea that individual persons are influenced by the social groups or contexts to which they belong (and vice versa) [84, 125, 187]. For example, self-rated health may relate to the population density of neighborhoods [125] and students' performance may depend on the skills of their teacher [84]. The individuals and social groups are conceptualized as a hierarchical system of individuals nested in groups, and groups nested in larger groups. The shared context introduces a correlation structure between individuals belonging to that context and the assumption that data is IID is violated.

Figure 3.2 gives a schematic illustration of such a dataset. Here, values of attributes a_{11} and ℓ_{11} are nested in the values of attributes a_{21} and ℓ_{21}, which again are nested in the values of attributes a_{31} and ℓ_{31}. Figure 3.2 thus depicts hierarchical data with three levels.

In hierarchical data, the lowest level is not necessarily that of the individual. As a real-world example of Figure 3.2, consider a series of blood glucose measurements sampled from $n = 5$ patients. Patients $i \in \{1,2,3\}$ are treated by one doctor, whereas patients $i \in$

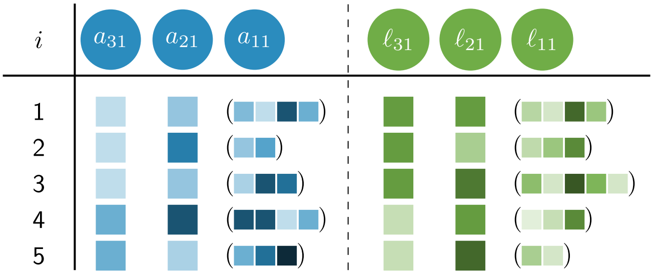


Figure 3.2: Schematic illustration of a hierarchical dataset with three levels. Attributes a_{11} and ℓ_{11} reside at the lowest hierarchical level, attributes a_{21} and ℓ_{21} reside at the second level and attributes a_{31} and ℓ_{31} reside at the highest hierarchical level. In the context of EMM, we distinguish descriptive attributes (blue) from target attributes (green).

$\{4,5\}$ are treated by another doctor. We may denote the blood glucose measurements of patient $i \in \{1,2,\ldots,n\}$ as $\mathbf{z}^i = (z^{i1}, z^{i2}, \ldots, z^{i\,t_z})$, where $t \in \{1,2,\ldots,t_z\}$ is the t^{th} hospital visit. In Figure 3.2, this is represented as one tuple for row i in column ℓ_{11} (e.g., $t_z^1 = 4$). Furthermore, patients may be described by a feature vector $\mathbf{x}^i = \left(x_1^i, x_2^i, \ldots, x_d^i\right)$. In Figure 3.2, only the first value of this feature vector is shown as a_{21}^i. The other values would appear under attributes $a_{22}, a_{23}, \ldots$ and so on.

Often, we consider n individuals to be IID, that is, $p\left(\mathbf{X}^1,\ldots,\mathbf{X}^n\right) = \prod_{i=1}^n p\left(\mathbf{X}^i\right)$. However, since the blood glucose values are nested, the $t_z^1 + t_z^2 + \ldots + t_z^n$ blood glucose values should not be considered independent, that is, $p\left(\mathbf{Z}^1,\ldots,\mathbf{Z}^n\right) \neq \prod_{i=1}^n p\left(\mathbf{Z}^i\right)$. If the observations at the lower hierarchical level represent independently drawn entities, we can assume that the nested measurements are IID *given the observations of the higher level entity*. This applies, for instance, when students are nested in school classes, or when survey respondents are nested in neighborhoods.

In the context of repeated measurements. we could theoretically assume that there is no effect of time (i.e., the blood glucose values do not depend on the measurement occasion). Then, we may assume the nested measurements are IID *given an individual patient*. We write $p\left(\mathbf{Z}^1,\mathbf{Z}^2,\ldots,\mathbf{Z}^{t_z} \mid \mathbf{X}\right) = \prod_{t=1}^{t_z} p\left(\mathbf{Z}^t \mid \mathbf{X} = \mathbf{x}^i\right)$. However, it is likely that for time-varying data, such an assumption does not hold.

3.3 Related work

Several authors have provided unified terminologies and hierarchies for DM. At the second-ever edition of KDD, [58] distinguished Knowledge Discovery in Databases (KDD) from DM. Here, KDD would be the “overall process of discovering useful knowledge from data, while data mining refers to a particular step in this process” [58, p.82]. In this distinction, data selection, data preparation, and appropriate interpretation of the results, are all crucial steps in the KDD process.

Zooming in on pattern mining, an umbrella paper on Supervised Descriptive Rule Discovery [110] unifies contrast set mining (CSM), emerging pattern mining (EPM) and SD.

Zooming in further, surveys on SD are available [8, 76, 80]; regarding EMM, the paper that comes closest to a survey paper is [54]. Since these works were published, both SD and EMM have further developed, especially towards the discovery of subgroups in datasets that have some form of hierarchical structure. In this chapter, we provide a unified terminology for EMM for such type of work.

In this chapter, we build on terminology from multiple domains. As introduced before, we build on terminology in multilevel analysis (MLA), which is particularly well-known in the biomedical and social sciences and adopts the concept that individual persons are influenced by the social groups or context to which they belong (and vice versa) [84, 125, 187]. The individuals and social groups are conceptualized as a hierarchical system of individuals nested in groups, and groups nested in larger groups. Since RVs may be defined at any hierarchical level, a dataset has a hierarchical or nested structure [84].

Furthermore, we build on terminology in relational databases (RDBs). An RDB is a collection of row-by-column tables (or, relations) where each table contains information on one entity type. Rows are also called tuples, and columns are attributes. Originally and most commonly, RDBs follow the First Normal Form (1NF) that no attribute domain has relations as its elements [33]. Hence, we refer to single-table data (cf. Figure 2.1) as an IID dataset. Furthermore, an entity-relationship model (ER) for an RDB classifies the things of interest as entity types, and specifies which relationships exist between entities (instances of the types) [31].

In this dissertation, we consider data to have a hierarchical or nested structure if the entities have one-to-many relationships. Note that the distinction between one-to-many and many-to-many relationships is not necessarily determined by the data itself, but also by the domain-specific interpretation and definitions of the entities, entity types and the attributes. In our unified terminology, we aim to classify existing work based on the explanation of the relation between concepts provided by the authors.

Our unified terminology demonstrates that EMM for hierarchical data requires some form of data manipulation. We come across different approaches such as joining RDB tables in a long (stacked) flat-table format (Section 3.5.1) or applying some form of aggregation or propositionalization [60] (Section 3.5.3). Hierarchical data formatted as in Figure 3.2 has similarities with the idea of Nested Relations where the assumption of 1NF does not hold and relations may have attribute values that are relations themselves [166].

3.4 Our proposed unified terminology

Formal terminology in EMM is agnostic about the origin of the data and whether or not record r^i is independent from record r^j. The term *record* is deliberately abstract: depending on the application domain, they may be objects, patients, sensors or something else. Nevertheless, dataset Ω generally consists of n independent entities and is formatted as in Figure 2.1 where each row contains an entity. Then, subgroup descriptions based on selection conditions filter out entities and subgroups are independent of their complement.

We extend EMM definitions to include data with a hierarchical structure. We first define a hierarchy following [19]:

Definition 3.1 (Hierarchy cf. [19])**.** *A hierarchy (tree) $\mathscr{H}$ is a tuple $\mathscr{H} = (E, \leq, e_0)$ where:*

- *$E = \{e_0, e_1, \ldots, e_N\}$ is a set of items,*
- *$\leq$ is a partial order relation defined over this set,*
- *$\forall e \in E : e_0 \leq e$ (item e_0 is called the root of $\mathscr{H}$), and*
- *there is only one path from e_0 to any other item: $\forall e_i, e_j, e_k \in E : e_i \leq e_k \wedge e_j \leq e_k \implies e_i \leq e_j \vee e_j \leq e_i$.*

We then use this definition as the fundament on which we build a definition for what makes a dataset hierarchical. We define hierarchical data as a collection of measurements taken from a set of entities E:

Definition 3.2 (Hierarchical data)**.** *Consider a collection of measurements from a set of entities E. We refer to these measurements as hierarchical data when the entities can be organized as a hierarchy $\mathscr{H} = (E, \leq, e_0)$ such that each of the entities is associated with an entity type $c_d \in \mathscr{C} = \{c_1, c_2, \ldots, c_C\}$, entities at the same level have the same entity type, and the depth of the hierarchy $\mathscr{H}$ is C. The leaf entities are associated with entity type c_1, this is the lowest hierarchical level. The measurements at level d are nested in the entities of entity type c_{d+1}. The root e_0 is the population and is not associated with an entity type.*

Figure 3.3 gives an example of hierarchical data formatted as a long table. Figure 3.4 displays its associated hierarchy $\mathscr{H}$. There are three entity types $\mathscr{C} = \{c_1 = t, c_2 = i, c_3 = h\}$. As an example, consider repeated measurements nested in patients, nested in hospitals. The lowest level does not necessarily have to be time. Consider an example where $\mathscr{C} = \{\text{person}, \text{household}, \text{neighborhood}\}$ or $\mathscr{C} = \{\text{student}, \text{program}, \text{university}\}$.[4] In Figures 3.3 and 3.4, entities e_1 and e_2 are of type h, entities $e_3, ..., e_7$ are of type i and entities $e_8, ..., e_{23}$ are of type t. If h, i and t are used as indicators, it is more common to start counting from 1 at every level ($h \in \{1,2\}$, $i \in \{1,2,3\}$ and $t \in \{1,2,3,4\}$, see Figure 3.3).

Each entity type $c_d \in \mathscr{C}$ has an associated set of RVs. In the context of EMM, we distinguish descriptive RVs from target RVs. The set of descriptors is $\mathbf{A} = \{\mathbf{A}^{(d)}\}_{d \in \{1,2,\ldots,C\}}$, where $\mathbf{A}^{(d)}$ are the descriptors at level d. Not all levels need necessarily have descriptors attached to them, but an RV describes only one entity type. To be consistent with EMM terminology, we refer to the observed measurements as attributes rather than RVs. We then write a_{dj} to refer to the j^{th} attribute measured at level d, for $d \in \{1, 2, \ldots, C\}$ and $j \in \{1, 2, \ldots, k_d\}$, where k_d is the number of descriptors at level d. For instance, attribute a_{31} (a_{23}) is the first (third) attribute to measure information of entities belonging to entity type $c_3 = h$ ($c_2 = i$). The same notation applies to target attributes.

For EMM with hierarchical data, one must be certain and clear about which entities should be grouped together in candidate subgroups. If there are multiple entity types, subgroups

[4] The distinction between entity types and categorical variables is not well defined. In general, we say that entities form a new hierarchical level if they can be *sampled* [125]. For instance, we can sample neighborhoods, schools and hospitals but not ethnic groups or genders. The latter are attributes associated to entities of the entity type individual. In some scenarios using a categorical variable to define a new hierarchical level can be considered legitimate, since such an approach closely connects to stratified sampling. In this dissertation, we assume such a scenario in Section 6.7: survey respondents are nested in nations.

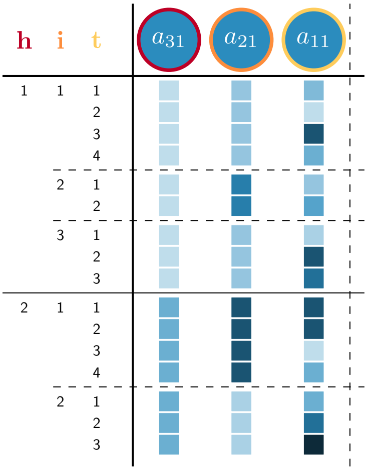


Figure 3.3: Hierarchical data formatted as a long (stacked) table. Associated hierarchy $\mathscr{H}$ is displayed in Figure 3.4.

of entities can be formed for any of these types. For instance, we may be interested in discovering exceptional behavior of hospital visits (e.g., visits with extremely low lab values), patients (e.g., patients with below average treatment effects) or hospitals (e.g., hospitals with high mortality rates). We propose the notion of *subgroup level*, the hierarchical level of the entity type for which subgroups should be formed:

Definition 3.3 (Subgroup level)**.** *Consider hierarchical data. The entity types are* $\mathscr{C} = \{c_1, c_2, \ldots, c_C\}$. *The subgroup level is the hierarchical level* d *of the entity type* $c_d \in \mathscr{C}$ *for which we aim to discover exceptional subgroups of entities. The number of entities at the subgroup level is denoted with* n.

Almost all work on EMM uses “conjunctions of selection conditions on descriptors **A**” as description language $\mathscr{D}$. In EMM with hierarchical data, data formatting choices interact with this description language. For instance, in the long format of Figure 3.3, the number of rows equals the number of entities at level c_1: the rows are associated with lowest-level entities. Consequently, a selection condition on attribute a_{31}, as traditionally used in EMM, will filter out single measurements. In contrast, a selection condition on attribute a_{21} (which describes entities at a higher hierarchical level than the level that corresponds to the rows) selects all rows belonging to the covered entities at level c_2. That is, the complete sub-trees of the covered entities are selected.

One could re-format the data such that every row contains information on entities of type c_2, as depicted in Figure 3.2. Here, the representation of attributes a_{31} and ℓ_{31} is with tuples rather than single values. In this format, a selection condition on attribute a_{21} filters out patients as single rows (rather than as groups of rows as before in Figure 3.3). In other words, the format in Figure 3.2 can be used to discover subgroups at level c_2, while the long flat-table format works better for discovering subgroups at level c_1.

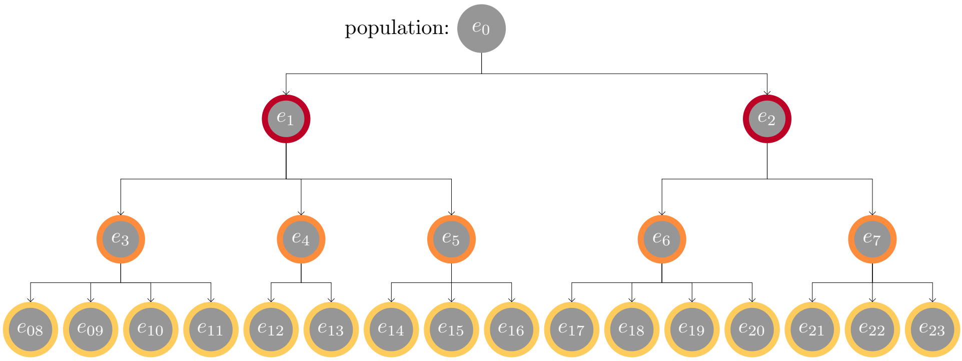


Figure 3.4: Hierarchy of the entities in the schematic datasets presented in Figures 3.2 and 3.3. The dataset contains three entity types $\mathscr{C} = \{c_1 = t, c_2 = i, c_3 = h\}$. Entities e_1 and e_2 are associated with entity type $c_3 = h$, entities $e_3, ..., e_7$ are associated with type $c_2 = i$ and entities $e_8, ..., e_{23}$ are of type $c_1 = t$.

In our unified terminology, we always format hierarchical data as in Figure 3.2. We ensure that the rows represent the entities of the chosen subgroup level (cf. Definition 3.3). Subsequently, we determine whether the existing descriptive and target attributes reside at a lower, the same, or a higher hierarchical level than the subgroup level.

Definition 3.4 (Lower level attribute)**.** *Consider hierarchical data with entity types* $\mathscr{C} = \{c_1, c_2, \ldots, c_C\}$ *and a pre-defined subgroup level* d^* *of entity type* c_{d^*}*. A descriptive attribute* a_{dj} *or target attribute* ℓ_{dj} *is a lower level attribute if* $d < d^*$ *for* $d, d^* \in \{1, 2, \ldots, C\}$.

Definition 3.5 (Same level attribute)**.** *Consider hierarchical data with entity types* $\mathscr{C} = \{c_1, c_2, \ldots, c_C\}$ *and a pre-defined subgroup level* d^* *of entity type* c_{d^*}*. A descriptive attribute* a_{dj} *or target attribute* ℓ_{dj} *is a same level attribute if* $d = d^*$ *for* $d, d^* \in \{1, 2, \ldots, C\}$.

Definition 3.6 (Higher level attribute)**.** *Consider hierarchical data with entity types* $\mathscr{C} = \{c_1, c_2, \ldots, c_C\}$ *and a pre-defined subgroup level* d^* *of entity type* c_{d^*}*. A descriptive attribute* a_{dj} *or target attribute* ℓ_{dj} *is a higher level attribute if* $d > d^*$ *for* $d, d^* \in \{1, 2, \ldots, C\}$.

For example, in Figure 3.2, the subgroup level is c_2. Then, attribute a_{31} resides at a higher hierarchical level, a_{21} at the same level and a_{11} at a lower hierarchical level than the subgroup level.

3.5 Classification of existing literature

We categorize existing EMM literature into 9 boxes based on whether the descriptive and target attributes reside at a lower, the same, or a higher hierarchical level (see Table 3.1). We believe such a categorization is needed because current literature on EMM for non-IID data use a mixture of notation, terminology, and re-formatting and pre-processing solutions. This obscures an overview of existing approaches, preventing comparison of approaches and identification of research gaps.

Table 3.1: Categorization of existing EMM methodologies. Descriptors and targets can be measured at a lower, the same or a higher hierarchical level than the subgroup level (which is given).

		Targets		
		lower	***same***	***higher***
Descriptors	***lower***	**A** [88, 93, 129]	**B** [66, 77, 103, 114, 209, 211] [57, 132] [131, 138, 159]	**C**
	same	**D** [42, 72, 156, 204] [26, 143] D^* [9, 20, 21, 47, 98] D^* [95]	**E** [17, 51, 54, 56, 109, 121] E_1^* [10, 50, 124] E_2^* [29, 85, 86, 106]	**F** [158] F^* [19]
	higher	**G** G^* [185]	**H** [118, 119, 140] [16, 147, 202, 203]	**I**

Methodology

We systematically search for existing work on EMM with non-IID data. We select relevant papers in the following order: 1) we are already familiar with the work and know it relates to SD or EMM (78 papers), 2) they are cited by work that we assigned to Boxes A-I (but not E) (+13 papers), 3) they cite work that we assigned to Boxes A-I (but not E) (+16 papers) and finally, 4) they cite [54] (+6 papers). Next, we translate the paper using our unified terminology; that is, we determine whether the work uses hierarchical data, what is the subgroup level, and whether the attributes reside at lower, the same, or higher hierarchical levels. While categorizing, we focus on information provided in the notation and terminology sections of the paper, the explanation of the descriptors, the explanation of the target model and if still unclear, the description of the datasets. Categorization of papers was done by two authors.

Next, we discuss the boxes one by one in the following order: E (descriptors and targets reside at the subgroup level) → D, G (lower level targets) → F (higher level targets) → B, A (lower level descriptors) → H (higher level descriptors). We provide additional discussion for papers that we assign to a box, but whose box membership is not as clear cut as it is for the other papers; such assignments are marked by an asterisk.

3.5.1 Box E: descriptors and targets reside at the subgroup level

All instances of SD and EMM that were developed for IID data can be assigned to category E in our unified terminology: the subgroup level equals the level of the descriptors and the targets. Essentially, IID data has only one entity type, and all attributes are measured for that particular entity type. For instance, [121] use the Windsor housing data to discover subgroups of houses with an exceptional relation between the lot size and sale prices and [54] analyze the association between race and completion of a vaccine regimen. The IID nature of datasets is not always directly clear. For instance, [54] describe a Gene dataset that shows the expression level of 313 genes for 63 patients. A first glance, the dataset may appear nested (i.e., genes nested in patients), but close reading reveals that the genes' expression levels are the descriptive attributes.

An interesting edge case is [17], who analyze the European Parliament Voting (EPV) dataset which contains information on a set of parliamentarians, a set of ballots, and the outcomes of votes of certain parliamentarians for certain ballots. Following RDB terminology, this dataset contains tables of two entity types with a many-to-many relationship. In practice, [17] search the two tables separately, discovering descriptions of parliamentarians (called, a group) and descriptions of ballots (called, contexts). They match these descriptions such that the contextual intra-agreement of a group is exceptional with respect to the non-contextual, overall agreement of that same group. We therefore classify this approach in Box E.

Overall, we have three reasons for arguing that most work on SD and EMM use IID data:

1. The proposed target models compare a subgroup to its complement using statistical quality measures that assume independence of observations,
2. The description language filters out rows, and the definition of a row is often equal to the definition of an individual (e.g., subgroup of patients, where every row is a patient), and
3. Many of the experiments use datasets with an IID nature.

Since Box E is the default box, we will not exhaustively enumerate all SD and EMM papers belonging to this box.

Box E^*: Reformatting hierarchical data into flat-table data

Some papers employ hierarchical data but belong to Box E anyway. These are the papers that format the hierarchical data in a long flat-table format (as in Figure 3.3) and use descriptors and targets that reside at the lowest hierarchical level, that is $\mathbf{A} = \mathbf{A}^{(1)}$ and $\mathbf{L} = \mathbf{L}^{(1)}$, respectively. Within this category, we can distinguish between two types of work.

On the one hand, some papers present their data as a row-by-column flat table but do not explicitly mention whether rows are considered independent. Astute readers may consider some of the experimental data to be hierarchical. In that case, there is an increased risk of unintentionally finding nested subgroups (cf. the example in Figure 3.1). For instance, [10] analyze real-world data from the WideNoise smartphone app where each data point includes objectively measured noise in decibel and an associated subjective perception of the noise plus a set of tags to provide semantic context, both given by the user. The risk is that the same user provides multiple data points and as a consequence, the description of discovered subgroups may not be as representative as solicited. Other examples are [50], who aim to discover subgroups of tweets (posts) with exceptional spatio-temporal behavior using a Bayesian non-parametric model that assumes the posts to be independent, and [106], who analyze exceptional claiming behavior in pharmacies by assessing claims (lowest-level) of one pharmacy relative to many other pharmacies (higher-level).

On the other hand, some papers explicitly mention the nested structure of their data. Eventually, after the data preparation, the subgroup level is at the lowest hierarchical level. An example is [124], who analyze sequences of states and introduce 1$^{\text{st}}$ order Markov Chains (MCs) as a model class. The authors specifically mention that they change the subgroup level: "we split the given state sequences in order to construct a tabular dataset, in which each instance corresponds to a single transition" [124, p.712].

Another example is [85], who convert time series into slices, and summarize each slice using features such as descriptive statistics and measures of complexity, the number of peaks or extreme points and auto-correlation. Some of these characteristics are used in descriptive space; others in target space. Specifically, the time series originate from speech and video recordings of 27 meetings, each consisting of 3-4 participants. Hence, the data has a clear hierarchical structure, but the exceptional subgroups contain the lowest-level entities: the slices. This allows the authors to discover patterns such as "low amounts of movement might be indicative of complex speech dynamics" [85, p.11]. At the same time, interpreting subgroup descriptions at a higher hierarchical level than the subgroup level is non-trivial; [86] use visualization techniques to do so.

3.5.2 Hierarchical target attributes

Box D: Lower level targets

The target attributes in an EMM instance may contain measurements at a lower hierarchical level than the subgroup level. For instance, [26] use a Dynamic Bayesian Network (DBN) containing three temporal RVs in target space, to discover exceptional subgroups of applications submitted to the European Union. Each application is considered independent (with descriptors such as land area, department, year of submission). In target space, each application has a series of workflow activities (i.e., events) related to it. Hence, the entity type event is nested in the entity type application. The work by [26] differs from [56] (which is assigned to Box E), since the nodes of the BNs used in [56] represent single RVs, whereas the variables in a DBN are repeatedly measured over time.

From a different perspective, [26] aggregate from a lower hierarchical level to the subgroup level by first, selecting all target data of the entities that are covered by the subgroup description and second, evaluating the parameters of a model estimated on those lower-level measurements. In contrast, [143] first estimate a mixed effects model per subgroup-level entity (using the lower-level target attributes) and then discover subgroups of exceptional parameter estimates. All target models discussed so far (DBNs, MCs, mixed effect models) are particularly well-suited for analyzing hierarchical data.

A series of work discuss Exceptional Preferences Mining (EPM) [42, 156, 204]; an approach for discovering subgroups of entities (districts) with an exceptional label ranking (the order of parties after elections). The party ranking is nested in the districts. A similar study was presented by [72], who additionally use collected statistics from survey data to capture the descriptive attributes of each district. Although the statistics are aggregations based on measurements from lower-level entities (persons within the districts), we assign [72] to Box D since those lower-level values were likely not available for analysis.

Box D^*: Exceptional subgraph mining

Mining exceptional subgraphs belongs in Box D, but often in a nontrivial manner. For instance, [9, 20, 21, 47] discover exceptional subgraphs using descriptive attributes of the vertices; [98] use attributed edges. We classify these works separately since 1) they only use descriptive attributes and no target attributes (exceptionality of subgraphs is evaluated using quality measures based on e.g. density, community-detection algorithms, modularity, Weighted Relative Accuracy (WRAcc)), and 2) a graph structure is non-IID but not necessarily hierarchical. With a bit of creativity, we could consider a directed graph to be a collection of measurements on entities of two entity types: vertices and edges, where measurements on edges are nested in vertices. Under certain constraints, we could do a similar thing for undirected graphs. All in all, exceptional subgraph mining mostly analyze datasets where each node is an independent individual, subgroup descriptions cover nodes, and the target model is a non-IID structure nested in the nodes. We therefore classify this category of work as D^*. Remark that our selection of papers is non-exhaustive; exceptional subgraph mining is a large sub-domain of EMM.

In the same sub-category, we explicitly mention [95], who have a “dataset of entities, each of which contains an arbitrary structure”. In descriptive space, entities are independent and selected using descriptive attributes. In target space, [95] propose a method to discover subgroups with deviating structure as either a clustering in the given reproducing kernel Hilbert space, or the most anomalous subgroup within that space. Their method is generic, and can be applied to entities with different structure, such as a graph, time series, or molecules.

Box G^*: Describing communities using ontologies

Furthermore, [185] discover exceptional subgraphs in a 2-step procedure. First, they discover communities in a knowledge graph and second, they describe these communities using ontologies (which can be considered a higher level descriptor). Therefore, we classify [185] in Box G^*. In Section 3.5.3, we will discuss using ontologies in descriptive space in more detail.

Box F: Higher level targets

Recently, [158] proposed mining Java memory errors using EMM with hierarchical targets. Specifically, the authors analyze OutOfMemory errors (each error is an independent entity) by analyzing how memory in a Java Virtual Machine is divided over a set of classes. These classes are organized hierarchically in packages. For instance, package `J.L.reflect.Method` is nested in `J.L.reflect.*`, which is nested in `J.L.*`). In our unified terminology, the packages are the set of target attributes $\mathbf{L} = \{\mathbf{L}^{(1)}, \mathbf{L}^{(2)}, \ldots\}$, where every RV L is a counter (i.e., integer type). We consider the entities of interest to be at the lowest hierarchical level (since we know the memory of the lowest-level packages) and targets reside at higher hierarchical levels: [158] is assigned to Box F.

Box F^*: Higher level targets without descriptors

The work by [19] leads [158] but does not use descriptors.

3.5.3 Hierarchical descriptive attributes

Box B: Lower level descriptors

Interestingly, most existing literature considering lower-level descriptive attributes are SD papers that assume one categorical target attribute and use a quality measure related to Weighted Relative Accuracy (WRAcc). In fact, a few of the first papers on SD address the discovery of subgroups in multi-relational data [103, 209]. To that end, [209, p.80] first define "a designated object relation intended to be the master relation of the population of interest" (similar to our notion of subgroup level as in Definition 3.3) and then use a sampling technique to filter the appropriate target data without joining all possible relations. Alternatively, [103] traverse the search space by using a set of aggregation functions such as count, average, max, and min, additionally exploiting a pre-defined relation graph that determines which joins are selected. Both [209] and [103] do not actively change the format of the data; that is, they work with the RDB structure.

Transforming an RDB into a single-data-table representation by way of propositionalization leads to Relational SD (RSD) [211]. Propositionalization is a flattening approach that constructs features at the level of the individual (i.e., the subgroup level). Specifically, [211] use Prolog queries consisting of structural predicates which refer to parts (substructures) of the entities of interest, and create a binary table where each column represents a newly created feature that may or may not be present for a particular entity. The binary table can then be searched using traditional search algorithms.

Some nested data originates when data from multiple sources are combined. For instance, [77] use a variety of discretization and aggregation functions in order to combine students' enrollment data with data obtained from the app Moodle, which describes students' online participation in activities and resources for various courses. Other examples where data from different sources are combined are [66], who gather data on various crises and use a set of time-varying descriptors that preceded the occurrence of each crisis. A similar approach was used in [114] for analyzing failed states.

So far, we have discussed two options for transforming a nested dataset with lower level descriptors into a flat-table data: 1) transformation into a long format by changing the subgroup level, work is placed in Box E^*, and 2) using aggregation functions to summarize the nested measurements (in descriptive space in Box B, in both descriptive and target space in Box A, still to be discussed). An alternative is to transform hierarchical data into a wide flat-table data format. This was proposed by [132], who developed an anytime algorithm for numerical time-varying descriptors and suggest dealing with time series by composing "a set of variables X at each timestamp in T. [...] each numerical attribute gives the value of a variable at a given time" [132, p.6].

Another approach is taken by [57], who use EMM in a Process Mining (PM) context to discover interesting subgroups of cases in help desk event logs. Specifically, [57] transform a *long* flat-table of event logs (entity type *log*) into a *wide* flat-table format where rows represent entities of entity type *case*. Then, the tuple measured for case i for an attribute a_{dj}^{it} is represented by $t_{i_{dj}}$ new attributes.

An interesting approach to handling lower level descriptors is taken by [131, 138], who adopt a description language other than a conjunction of selection conditions. Specifically, [131] and [138] discover exceptional subsequences in a dataset containing n independent objects encompassing a sequence and a label. As is common in sequential data mining, the subsequences discovered by [131, 138] are ordered lists of itemsets. Hence, the coverage of an object does not depend on the strict *location* of an itemset in the sequence, but rather on a similar *order* of itemsets. An application of the work of [131, 138] to PM is presented by [159].

Box A: Lower level attributes in both descriptive and target space

We found few instances of EMM that work with lower level attributes in both descriptive and target space. First, [93] discovers subgroups of households by analyzing the electricity consumption in kWh of every one of nearly 5 000 households, with 30-min resolution, from April 2009 to October 2010. Specifically, [93] construct consumption-related attributes such as the averaged maximal daily consumption and the absolute difference between the measured consumption and a profile estimate. These are then combined with household-specific variables (descriptors at the subgroup level). By using different aggregation approaches, [93] utilize the same sequence of electricity consumption in both descriptive and target space.

Furthermore, [88] use SD to create interpretable explanations of top-N recommendations made by a state-of-the-art recommender system. In target space, the N recommendations are averaged into one numerical value per user. In descriptive space, the user's past actions are given by sequences of items. The work by [129], who focus on top-1 recommendations, leads [88] In descriptive space, both [88, 129] perform SD for each individual user separately: the top-q results list of exceptional patterns explain the recommendations provided to an individual user. To this end, the authors generate hypothetical target values for perturbed sequences. Furthermore, they provide two types of exceptional patterns (explanations), each with a different description language. On the one hand, conjunctions of attribute-value pairs with binary attributes describe which actions (items) strongly impact the recommendations. On the other hand, exceptional subsequences cf. [131, 138] (Box B) explain whether the order of the items is important.

Box H: Higher level descriptors

Instead of the entities in the dataset, the attributes can find themselves in the hierarchical structure. We already discussed [158] who consider a hierarchy of target attributes (Box F). Here, we focus on [118, 119, 140, 202], who discover exceptional subgroups using semantic information in descriptive space. First, [118, 119] use RSD [211, itself part of Box B] to enable background knowledge in the form of ontologies to be used in relational data mining, called semantic SD or Semantic Data Mining. An ontology is a hierarchy of concepts, such as the nesting of cities in regions in countries [202], or the nesting of protein binding sites in genes [118, 119, 140] The hierarchical structure of the concepts is generally given by domain experts and every entity in the dataset has an associated value for each of the concept types. Consequently, ontologies are valuable for traversing the search lattice efficiently. For instance, a candidate description *Region* = *Bavaria* automatically selects all entities that reside in all cities of Bavaria.

Semantic information is also employed to discover exceptional subgroups in [16, 203]. Specifically, [203] apply traditional SD to find an initial set of subgroups, and then use ontologies to improve the descriptions, whereas [16] construct a Hierarchical Multi-tag Attribute (HMT) out of a set of tags associated with each entity at the subgroup level. For instance, a ballot *4.10.04-Gender equality* in the EPV dataset is nested in the collection of ballots with tag *4.10-Social policy*, which is nested in tag *4-Economic, social and territorial cohesion*. Hence, a restriction over an HMT attribute creates a set of tags, and then covers the objects that explicitly or implicitly contain one of the tags.

In addition to using existing ontologies for describing subgroups, [147] propose a method for creating a data-driven hierarchy of numerical intervals out of one or more numerical descriptors. Depending on the adopted search strategy such an approach may be valuable but other search algorithms may not require pre-discretization beforehand [135].

3.6 Research gaps

We present a unified terminology for EMM for hierarchical data. Our framework includes existing work that discovers exceptional subgroups in attributed graphs, sequential data, relational databases and in time series data. More work can be done. For instance, we discussed work that use ontologies and hierarchical attributes to traverse the search lattice efficiently. It would be valuable to investigate to what extent the proposed search algorithms can be generalized to include multiple ontologies and hierarchical attributes and whether those attributes can be combined with traditional attribute types.

We found few papers that address hierarchy in both descriptive and target space. Remark as well that Boxes C, G and I are still completely empty; Box F contains only one existing work. Possibly, EMM could be extended to other target models that handle hierarchical data such as location-scale models, Bayesian hierarchical models and ARIMA models.

Another interesting next step is to develop EMM to other types of non-IID and unstructured data such as image and text data. Such a research direction most likely includes the application of dimensionality reduction, representation learning and (automated) generation of (latent) attributes in descriptive or target space. The recent kernel-based method proposed by [95] could serve as a starting point.

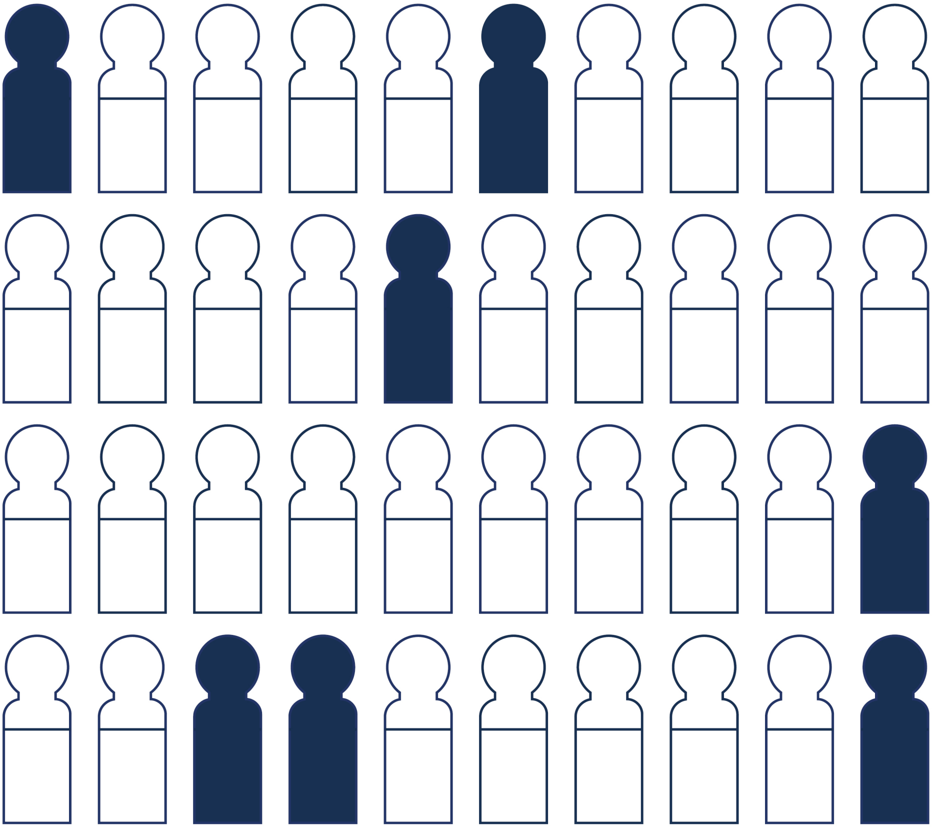
Lower level descriptors are most commonly handled by applying some kind of aggregation. Usually, these aggregation functions are created based on domain-specific knowledge; they are not generic. It would be interesting to explore the possibility to create aggregations in a more automated, data-driven way, possibly based on the evaluation of candidate subgroups in target space.

Lastly, we found few papers with a description language that differs from the typical conjunction of selection conditions. It would be valuable to investigate if new description languages, possibly in combination with other refinement operators, could be beneficial in further developing EMM towards non-IID data.

3.7 Conclusion

We provide a unified terminology for EMM with hierarchical data by 1) defining hierarchical data as a collection of measurements taken from different type of entities, where the measurements of one type of entity are nested in another type of entity, 2) the notion of a subgroup level as the hierarchical level of the entity type for which subgroups should be formed and 3) classifying existing literature into 3 × 3 classes based on whether descriptive and target attributes reside at lower, the same, or higher hierarchical levels. Naturally, we have aimed at letting our classification cover all relevant work on SD and EMM, with a special focus on non-IID data; there exists no guarantee that we may not have missed something. Nevertheless, our unified terminology provides insight into many interesting, existing approaches for EMM with hierarchical data and demonstrates valuable and promising future directions.

4



Mining Exceptional Transition Behavior of Varying Order

Discrete Markov chains are frequently used to analyze transition behavior in sequential data. Here, the transition probabilities can be estimated using varying order Markov chains, where order k specifies the length of the sequence history that is used to model these probabilities. Generally, such a model is fitted to the entire dataset, but in practice it is likely that some heterogeneity in the data exists and that some sequences would be better modeled with alternative parameter values, or with a Markov chain of a different order. We use the framework of Exceptional Model Mining (EMM) to discover these exceptionally behaving sequences. In particular, we propose an EMM model class that allows for discovering subgroups with transition behavior of varying order. To that end, we propose three new quality measures based on information-theoretic scoring functions. Our findings from controlled experiments show that all three quality measures find exceptional transition behavior of varying order and are reasonably sensitive. The quality measure based on Akaike's Information Criterion (AIC) is most robust for the number of observations.

The contents of this chapter have previously appeared in **Schouten, R. M., Bueno, M. L., Duivesteijn, W., and Pechenizkiy, M.** Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery 36* (2022), 379–413 [170].

4.1 Introduction

Markov models in all their variants are frequently used to mine patterns in sequential data. Consider, for instance, 1^{st} order Markov chains [151, 165, 207], Hidden Markov Models (HMM) [27, 91, 136, 148] and Dynamic Bayesian Networks (DBN) [26, 37]. All these models are called *memoryless* if they satisfy the *Markov property*: given the data at time $t-1$, the data at time t is independent of the data before time $t-1$. Furthermore, a model is *homogeneous* if its parameters do not change over time, and the sequences are *stationary* if the initial values follow the same model [214].

We consider discrete Markov chains where the observations are discrete values, or states, from a countable set which is called the state-space. Generally, such a model is fitted to the entire dataset and the parameter estimates give information about the average transition behavior between states. However, some heterogeneity in the data often likely exists, and hence some sequences would be better modeled separately. We use Exceptional Model Mining (EMM) [54, 121] to discover these exceptionally behaving sequences.

EMM is a local pattern mining technique seeking subsets of the dataset that behave somehow exceptionally. Here, exceptional behavior is measured in terms of parameters of a *model class* over target attributes. A *quality measure* quantifies this exceptionality (see Section 2.3). Since EMM allows for ≥ 2 attributes to be part of the target model, it can be seen as a generalization of Subgroup Discovery (SD) [80, 102, 209], which uses 1 target attribute. Both frameworks employ a rule-based description language where resulting subgroups are described as a conjunction of attribute-value pairs.

An EMM model class exists for 1^{st} order Markov chains [124]. We extend their work by considering Markov chains of varying order, where order k specifies the length of the sequence that is used as *memory* in the model. Specifically, our method allows for discovering subgroups in situations where the order of the Markov chain differs between the subgroup and the dataset. This situation requires comparing unequal numbers of parameters. Hence, we do not use a parameter-based quality measure, as is common in EMM, but show how information-theoretic scoring functions can evaluate a subgroup's exceptionality.

We furthermore add to existing work by seeking subgroups of sequences, as opposed to subgroups of transitions [124]. Whereas the latter detects heterogeneity within sequences, we find subgroups of homogeneous sequences that are heterogeneous w.r.t. the entire dataset. Our model class is practically relevant for identifying the originator of an exceptional sequence, such as a patient with exceptional blood glucose fluctuations (see Section 5.5) or an atypical user session in click-stream data [164].

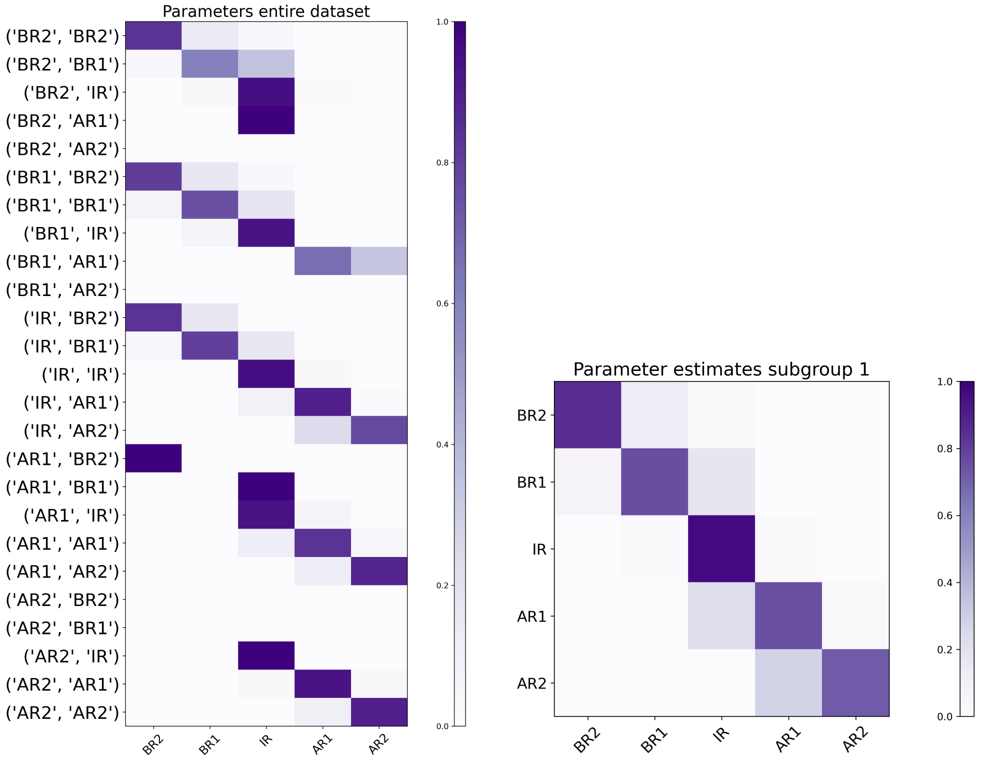
In sum, our main contributions include:

1. an EMM model class for detecting exceptional transition behavior of varying order,
2. a new set of quality measures based on information-theoretic scoring functions,
3. an understanding of how descriptive attributes can be used to form subgroups of entire sequences.

4.2 Background

The methods developed in this chapter are inspired by the second DIAbetes and LifEstyle Cohort Twente (DIALECT-2) [67]. DIALECT-2 is an observational study of adult persons with diabetes type 2, where blood glucose is measured every 15 minutes for a period of 14 days. Together with domain experts, we decide to discretize the continuous measurements into five blood glucose levels: below range 2 (BR_2), below range 1 (BR_1), in range (IR), above range 1 (AR_1), and above range 2 (AR_2), where range refers to desired blood glucose values. Although discretization may result in some information loss, these 5 blood glucose levels are the medical standard in treatment of diabetes type 2 [38]; they allow us to analyze transition patterns using discrete Markov chains.

As a running example to aid illustration of concepts introduced in the subsequent sections, some DIALECT-2 transition patterns can be found in Figure 4.1. In Sections 5.3 and 5.5, we discuss more details on the study and present two sets of results obtained by performing two analyses with our proposed methodology.



(a) Entire dataset.

(b) Subgroup of patients with *HbA1c category* = *low* ∧ *diabetes duration* ≤ *20 years* ∧ *21.3* ≤ *BMI* ≤ *35.7*, containing $N^{SG} = 22$ sequences (cov: 18%).

Figure 4.1: Transition patterns of blood glucose levels. (a) The entire dataset follows a 2nd order Markov chain. (b) The top subgroup follows a 1st order Markov model.

4.2.1 Data structure

Assume a dataset Ω with N independent but not identically distributed sequences of discrete random variables $X_t : t \in \{1, 2, ..., T\}$. The data realization at time t is denoted with x_t. Although sequence $r \in \Omega$ has length T_r, without loss of generality, we assume one fixed length T for every sequence. We refer to N as the data *size* and write M to denote the total number of *observations*, where $M = \sum_{r \in \Omega} T_r = NT$ (note that the total number of *transitions* is $M - N$). The set of possible discrete values is $V = \{v_1, v_2, ..., v_S\}$ for all x_t. For instance, in the DIALECT-2 dataset, $N = 126$, $T = 1344$, $M \approx 170000$, $V = \{\mathrm{BR}_1, \mathrm{BR}_2, \mathrm{IR}, \mathrm{AR}_1, \mathrm{AR}_2\}$ and $S = 5$.

We assume the availability of an extra set of attributes with information about the sequences. The are called the descriptive attributes. The full form of sequence r then becomes $(x_1, x_2, ..., x_{T_r}, a_1, a_2, ..., a_m)$ for all $r \in \Omega$. Here, m simply denotes the number of descriptive attributes. Depending on the application, these attributes could describe personal or medical characteristics such as HbA1c category, duration of the illness and BMI (Figure 4.1b), user session information such as browser language and timezone (if the sequences are click-streams) or contain meta-information about the sequences such as its length, state-space and starting time (see Section 4.6.1). We now explain how these descriptive attributes are used to form subgroups.

Exceptional Model Mining (EMM) is a local pattern mining framework, seeking subgroups in a population that behave somehow exceptionally [54, 121]. Those subgroups have interpretable descriptions and explainable circumstances under which exceptional behavior occurs.

Compared to the generic terminology for EMM, introduced in Section 2.3, in this chapter we define a subgroup such that a sequence is either covered by a description or not. We do not allow for a sequence to be split into pieces and to be partly assigned to a subgroup. The reason is that we want to find the originators of exceptional sequences because this could assist domain experts in adopting appropriate policies. For instance, an interpretable description of patients could assist doctors in selecting the most useful treatment; descriptions that only partly include certain patients are less helpful.

The strict partitioning between target and descriptive attributes is a powerful feature in EMM, allowing us to form subgroups of independently distributed individuals while analyzing sequential patterns. We are thus able to make subgroups of entire sequences because the descriptive attributes contain sequence-level information (see Section 4.2.1). In contrast, [124] find subgroups of transitions, using descriptive information on the transition (or time) level.

A quality measure quantifies the difference between behavior within the subgroup and behavior within the entire dataset (or the subgroup's complement) (see Definition 2.2). Generally, quality measures directly compare one or more parameter estimates, such as the difference between estimated slopes in a regression model [53] or the difference between estimated correlations of two target attributes [54]. Following the terminology of [188], we call these quality measures *parameter-based*. Their advantage is that you immediately know why a resulting subgroup is exceptional. However, a parameter-based approach also restricts the model in the subgroup to have the same number of parameters as the model in the entire dataset. In case of Markov chains, for instance, parameter-based quality measures would not allow the subgroup to be fitted with a higher (or lower) order model than the one that is fitted in the entire dataset. In this chapter, we therefore propose to evaluate a subgroup's exceptionality using quality measures based on information-theoretic scoring functions. We call these quality measures *evaluation-based*. For instance, in the DIALECT-2 study, the entire dataset is best modeled with a second-order Markov chain, as illustrated in Figure 4.1a, while the top subgroup is best modeled with a first-order Markov chain, as illustrated in Figure 4.1b. Our evaluation-based quality measures can gauge the exceptionality of the difference between these two models and their respective transition probabilities.

4.2.2 Markov chains

We will now first introduce 1st order Markov chains, and then extend the principles to k^{th} order chains. In this section, we overload the symbol Ω to refer only to the target attributes x_t for all $t \in \{1,2,...,T\}$ and temporarily forget about the descriptive attributes. We write SG to denote the same set of attributes for the subgroup.

Using the product rule and the Markov property that given the data at time $t-1$, the data at time t is independent of the data before time $t-1$, the joint probability distribution of Ω is modeled with a 1st order Markov chain by

$$P(\Omega|\theta) = P(\mathrm{x}_1,\mathrm{x}_2,...,\mathrm{x}_T|\pi,\mathbf{A}) = p(\mathrm{x}_1)\prod_{t=2}^{T} p(\mathrm{x}_t|\mathrm{x}_{t-1}). \tag{4.1}$$

The prior distribution $p(\mathrm{x}_1)$ is parameterized; it has an *initial probabilities vector* $\pi = [\pi_1,...,\pi_S]$. The main interest is in the transition behavior between time $t-1$ and t, which is parameterized with an $S \times S$ probability matrix denoted with $\mathbf{A}$. Parameters $\alpha_{ij} \in \mathbf{A}$ $\forall i,j \in \{1,2,...,S\}$ are estimated using Maximum Likelihood Estimation (MLE) where

$$p(\mathrm{x}_t = v_j|\mathrm{x}_{t-1} = v_i) = \alpha_{ij} = \frac{n_{ij}}{\sum_{j=1}^{S} n_{ij}}. \tag{4.2}$$

Here, n_{ij} denotes the total number of transitions from source state v_i to target state v_j $\forall i,j \in \{1,2,...,S\}$. Equation (4.2) thus practically means that we first calculate a transition frequency matrix, and then calculate the probabilities by dividing by the sum of each row. Consequently, $\forall_i \sum_{j=1}^{S} \alpha_{ij} = 1$.

Figure 4.1b shows such a 1st order transition probability matrix. The dark purple square in the top left corner expresses the probability (which is $\alpha_{11} = 0.85$) that the next blood glucose level is BR_2 (column) given that the current blood glucose level is BR_2 (row). The high probabilities on the diagonal indicate that patients are likely to stay at the same blood glucose level, although patients with a current blood glucose level of AR_2 are also quite likely to transition to a lower blood glucose value (to AR_1, $\alpha_{54} = 0.28$).

The Markov chain model in Equation (4.1) assumes *homogeneous* sequences where the transition parameters do not change over time [214]. If we additionally assume that the initial probabilities vector π follows that same transition model, we say the sequences are *stationary* [214]. It makes sense to make both assumptions together. After all, if we assume that the transition behavior does not change between time points 1 and T, it does not matter where or when the sequence starts. For example, if we estimate that 30% of the sequences move from state v_i to state v_j, it is also likely that 30% of the sequences start with state v_i. As a consequence, it is not necessary to separately estimate the parameters in π. Instead, we derive the initial probabilities by normalizing over all j target states in the frequency matrix. The total number of free parameters in a 1st order Markov chain is therefore $K = S(S-1)$.

However, depending on the application or for very short sequences, the starting point of the sequence could be of separate interest. Consider, for example, a subgroup of patients who present themselves with different symptoms than the overall patient population. In that case, the parameters in π are separately estimated using only the data from the first time point,

$$p(\mathrm{x}_1 = v_h) = \pi_h = \frac{n_h^{(t=1)}}{\sum_{h=1}^{S} n_h^{(t=1)}}. \tag{4.3}$$

Here, we indicate the selection of time points in the superscript $(t = 1)$. Separately estimating initial probabilities would add another $S-1$ free parameters to the Markov chain.

Extending the 1st order Markov chain model to a kth order model gives the following joint probability distribution,

$$P(\Omega|\theta) = P(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_T|\theta) = p(\mathrm{x}_1) \cdot p(\mathrm{x}_2|\mathrm{x}_1) \cdot ... \cdot p(\mathrm{x}_k|\mathrm{x}_{k-1}, ..., \mathrm{x}_2, \mathrm{x}_1) \cdot \prod_{t=k+1}^{T} p(\mathrm{x}_t|\mathrm{x}_{t-k}, \mathrm{x}_{t-k+1}, ..., \mathrm{x}_{t-2}, \mathrm{x}_{t-1}).$$

Such a model uses the memory of time $t-1, t-2, ..., t-k$ to predict the state value at time t. To understand the transition matrix, it can be helpful to consider the k-length history as one time point with S^k possible states. Transition matrix ${}_k\mathbf{A}$ is then an $S^k \times S$ probability matrix where value ${}_k\alpha_{ij}$ models the probability of moving towards state v_j $\forall j \in \{1, 2, ..., S\}$ given the $i \in \{1, 2, ..., S^k\}$ k-length history. Figure 4.1a shows the probability matrix of such a 2nd order Markov chain, where the rows represent the $5^2 = 25$ possibilities of a 2-length history given 5 blood glucose levels.

In higher order Markov chains, the main interest is still in the transition behavior and under the assumption of stationary sequences, the k initial probability distributions are often ignored. If needed, those initial probabilities can be calculated by normalizing over the last time point in the k-length history, just as we calculated π by normalizing over all j target states. We denote the normalization of ${}_k\mathbf{A}$ down to the ℓth order with a tilde: ${}_k^\ell\tilde{\mathbf{A}}$ $\forall \ell \in \{0, 1, 2, ..., k-1\}$. Its parameters are then written as ${}_k^\ell\tilde{\alpha}_{ij}$. For this reason, the number of free parameters in a kth order Markov chain is ${}_kK = S^k(S-1)$.

In Section 4.4, we will discuss quality measures based on information-theoretic scoring functions. These quality measures use log likelihood to quantify the goodness of fit of a given Markov chain. Here, we start making a distinction between two datasets: on the one hand, the dataset *for* which we calculate the goodness of fit, and on the other hand, the dataset *on* which we estimate the parameters of the Markov chain. We denote the latter with a superscript; ${}_k\mathbf{A}^F$ refers to a k$^{\text{th}}$ order transition matrix estimated on dataset F. We now calculate the log likelihood of dataset G using parameter estimates ${}_k\mathbf{A}^F$ by

$$\mathscr{L}(P(G|{}_k\mathbf{A}^F)) = \sum_{h=1}^{S} n_h^{G(t=1)} \log {}^0_k\tilde{\alpha}^F_h + \sum_{i=1}^{S}\sum_{j=1}^{S} n_{ij}^{G(t=\{1,2\})} \log {}^1_k\tilde{\alpha}^F_{ij} + \ldots$$
$$+ \sum_{i=1}^{S^{k-1}}\sum_{j=1}^{S} n_{ij}^{G(t=\{1,2,\ldots,k\})} \log {}^{k-1}_k\tilde{\alpha}^F_{ij} + \sum_{i=1}^{S^k}\sum_{j=1}^{S} n^G_{ij} \log {}_k\alpha^F_{ij}.$$

In the rest of this chapter we use $\mathscr{L}(P(SG|\mathbf{A}^{\Omega}))$ and $\mathscr{L}(P(SG|\mathbf{A}^{SG}))$ to denote the log likelihood score for a subgroup using parameter values estimated on the entire dataset and on the subgroup respectively.[1]

4.3 Related work

An EMM model class for 1$^{\text{st}}$ order Markov chains was introduced before [124]. There, the authors focus on finding subgroups of transitions and thus detect heterogeneity within sequences. We propose to extend this model class such that 1) we find subgroups of entire sequences and detect homogeneous sequences that are heterogeneous with respect to the other sequences, and 2) we allow for discovering subgroups that are best modeled with a different order Markov chain.

Since [124] considered the situation where subgroups follow the same 1$^{\text{st}}$ order model as the entire dataset, they proposed a parameter-based quality measure related to the *total variation distance* or Manhattan distance:

$$\omega_{tv}({}_1\mathbf{A}^{SG}, {}_1\mathbf{A}^{\Omega}) = \sum_{i=1}^{S}\left(\sum_{j=1}^{S} n_{ij}^{SG} \sum_{j=1}^{S}\left|{}_1a^{SG}_{ij} - {}_1a^{\Omega}_{ij}\right|\right). \tag{4.4}$$

The quality measure ω_{tv} can be extended to situations where the subgroup follows the same higher order Markov chain as the entire dataset, but it cannot be used in situations where the subgroup follows a different order model.

Model-Based Subgroup Discovery (MBSD) was proposed by [189, 190]. There, the divergence between the target probability estimates and the true labels of an outcome variable is evaluated using Proper Scoring Rules (PSR) [70]. We analyze sequential data without labels, but our evaluation measures are still related to those in [190] since the information-theoretic scoring function AIC is derived from the Kullback-Leibler divergence [28], which is associated with the logarithmic score as a PSR [70].

[1] Note that while calculating the log likelihood, we use normalized probabilities for the first k time points. In general, in this chapter we assume homogeneous and stationary sequences, except for Section 4.5.2, where we analyze non-stationary sequences.

In fact, for outcome variables with a probability density distribution, [188] defines a quality measure called *weighted divergence* where the information gain of the subgroup is calculated using log likelihood as the negative of the expected loss,

$$\varphi_{WD}(SG,\theta^{SG},\theta^{\Omega}) = \mathscr{L}(P(SG|\theta^{SG})) - \mathscr{L}(P(SG|\theta^{\Omega})). \tag{4.5}$$

Although our quality measures based on information-theoretic scoring functions only differ from φ_{WD} by the addition of a penalty for model complexity, it is exactly this penalty that allows for discovering subgroups with a different order Markov chain. In Section 4.5 we will show the difference in performance between our quality measures and φ_{WD}.

Several papers proposed information-theoretic scoring functions as the basis of a quality measure. For tabular data, [127] seek exceptional location and spread in multiple real-valued targets using a quality measure based on the information gain of subgroups that allows the user to incorporate prior knowledge in the process. For graph data, [47] aim to identify pairs of subgraphs with exceptional connectivity by looking at the density between the subgraphs, also allowing for the incorporation of prior knowledge.

In the context of sequence data, [26] consider dynamic Bayesian networks as model class and use BIC to define a mismatch score between the subgroup and its complement. We look at Markov chains, and compare the subgroup to the entire dataset because this is conceptually easier to understand and computationally more efficient than comparing to the subgroup's complement.

Like [26, 124, 188], we take an approach where candidate subgroups are evaluated using a bottom-up, heuristic search through the descriptive space. In comparison, [14] take a top-down approach where the subgroups are hypothesized beforehand based on theory and evaluated using Bayes Factors. [101] hypothesize two groups of sequences based on descriptive information and distributional characteristics. The two groups are analyzed with a Markov model and compared on their prediction accuracy.

A global approach to detecting (groups of) outliers is taken by [164], who calculate the log likelihood scores of individual sequences under a 1st order Markov model. Specifically, they create a Mahalanobis distribution of sequences by combining these log likelihood scores with meta information such as the sequence length. Although their approach points at unusual sequences in an existing dataset, it does not describe or explain in any other way why specifically those sequences are considered outliers. In contrast, the framework of EMM is a local pattern mining technique that not only allows for interpretable descriptions of exceptional subgroups but also for an explanation of why those subgroups are selected.

Yet, in Section 4.5, we will compare our method against a quality measure that only uses a globally fitted model and does not require the fit of separate models in each candidate subgroup. In particular, we compare against a quality measure called *weighted relative likelihood* [188],

$$\varphi_{WRL}(SG,\Omega,\theta^{\Omega}) = M^{SG} \cdot \left| \frac{\mathscr{L}(P(\Omega|\theta^{\Omega}))}{M^{\Omega}} - \frac{\mathscr{L}(P(SG|\theta^{\Omega}))}{M^{SG}} \right|. \tag{4.6}$$

Here, the average fit of the dataset is evaluated under parameters estimated on the dataset and compared against the average fit of the subgroup under those same parameters. Note that M^{SG} and M^{Ω} denote the total number of observations in the subgroup and dataset respectively and not the total number of sequences (see Section 4.2.1).

Recently, [138] propose Sequential Exceptional Pattern Discovery using Pattern-Growth (*SEPP*) as a general approach to the problem of Sequential Exceptional Model Mining (SEMM). Their work combines EMM with sequential pattern mining, the task to identify frequent subsequences, and as such they develop a search strategy based on GP-growth [123] and PrefixSpan [149]. [131] also combine sequential pattern mining with EMM, developing the *MCTSExtent* method building on Monte Carlo Tree Search (MCTS) [24]. Both MCTSExtent and SEPP consider the data to be in the traditional sequential pattern mining form where X_t is an itemset. Such data is inherently binary: an item is present in an itemset or not. A subgroup's exceptionality is then evaluated in terms of frequency or precision. In contrast, our sequences come with m descriptive attributes; subgroups are formed using these attributes and evaluated based on exceptional sequential behavior.

We now explain how we derive our quality measures based on information-theoretic scoring functions in Section 4.4. We also discuss the proposed search strategy.

4.4 Our proposed approach: Quality measures based on information-theoretic scoring functions

In order to evaluate the exceptionality of candidate subgroups with varying order Markov chains, we develop a set of quality measures that allows for the comparison of two sets of parameters of different size. Such quality measures should not simply select the subgroup with the largest number of parameters, because a more complex model may over fit the data. The quality measures should further take the subgroup size into account, since deviations from the norm are more easily obtained in smaller subgroups.

To that end, we base our quality measures on information-theoretic scoring functions. In general, for a dataset G, such as scoring function is defined by

$$\phi_{LL}(G,\theta^G) = \mathscr{L}(P(G|\theta^G)) - f(M^G)\cdot K^G, \tag{4.7}$$

where we use subscript LL to indicate that we use log likelihood as a way to quantify the goodness of fit. Given that θ^G are MLE parameters, $\mathscr{L}(P(G|\theta^G)$ is maximal. The second part of the equation is a penalty for model complexity. Here, K^G denotes the number of free parameters in a model estimated on dataset G. The term $f(M^G)$ is a penalty based on the number of observations in G.

We will apply three information-theoretic scoring functions; they differ in their penalty term. First, [3, 4, 28] derived that the bias in the log likelihood score (due to over fitting) converges to K as $M \to \infty$ (we temporarily leave out the superscript G). In Akaike's Information Criterion (AIC), the penalty is therefore set to K, which means that $f(M) = 1$. Second, the Bayesian Information Criterion (BIC) [180] sets $f(M) = \frac{1}{2}\log M$. In this chapter, we refer to BIC as an information-theoretic scoring function because it only differs from AIC by the extent of the penalty. However, BIC is derived from a Bayesian viewpoint, is related to Bayes factors [97] and originally focuses on model selection instead of prediction accuracy [152]. Third, a scoring function called AIC with small sample correction (AICc) penalizes with an additional $\frac{2K^2+2K}{M-K-1}$. The term corrects for over fitting if the number of free parameters is large with respect to M, but AICc converges to AIC as M increases [28, 87, 193].

Several authors have investigated the use of AIC and BIC in determining the appropriate Markov chain order [e.g., 168, 184, 195]. We use these scoring functions to evaluate whether candidate subgroups have exceptional transition behavior. This goes as follows.

In the situation that dataset Ω is heterogeneous and contains one or more subgroups of sequences that follow a different model than the rest of the sequences, it is likely that the parameters of the subgroup, θ^{SG}, describe the subgroup better than the parameters of the entire dataset, θ^{Ω}. This means that in the presence of a subgroup, the log likelihood of dataset Ω will increase if the parameters of the subgroup are separately estimated and evaluated. We write that

$$\mathscr{L}(P(SG|\theta^{SG})) + \mathscr{L}(P(SG^C|\theta^{\Omega})) > \mathscr{L}(P(\Omega|\theta^{\Omega})), \tag{4.8}$$

where SG^C denotes the subgroup's complement. Since $\mathscr{L}(P(SG^C|\theta^{\Omega}))$ is part of both the left and the right side of Equation (4.8), we can write that

$$\mathscr{L}(P(SG|\theta^{SG})) > \mathscr{L}(P(SG|\theta^{\Omega})). \tag{4.9}$$

We now derive our quality measures by combining Equations (4.7) and (4.9). We furthermore multiply ϕ_{LL} with -2 for conventional reasons, and again multiply with -1 to obtain quality measures that should be maximized (see Definition 2.2). This gives us the following three quality measures:

$$\varphi_{AIC} = 2\mathscr{L}(P(SG|\theta^{SG})) - 2K^{SG} - 2\mathscr{L}(P(SG|\theta^{\Omega})) + 2K^{\Omega}, \tag{4.10}$$

$$\varphi_{BIC} = 2\mathscr{L}(P(SG|\theta^{SG})) - K^{SG}\log M^{SG} - 2\mathscr{L}(P(SG|\theta^{\Omega})) + K^{\Omega}\log M^{SG}, \tag{4.11}$$

$$\begin{aligned}\varphi_{AICc} = 2\mathscr{L}(P(SG|\theta^{SG})) - 2K^{SG} - \frac{2K^{SG^2}+2K^{SG}}{M^{SG}-K^{SG}-1} - \\ 2\mathscr{L}(P(SG|\theta^{\Omega})) + 2K^{\Omega} + \frac{2K^{\Omega^2}+2K^{\Omega}}{M^{SG}-K^{\Omega}-1}.\end{aligned} \tag{4.12}$$

Note that quality measure φ_{WD} as defined earlier in Equation (4.5) sets $f(M) = 0$ and therefore uses no penalty. Furthermore, if the subgroup has the same order Markov chain model as the dataset, $K^{\Omega} = K^{SG}$ and the penalty terms cancel out for all three proposed quality measures.

4.4.1 Extended beam search algorithm

Beam search is a commonly used strategy to search through the space of candidate subgroups. It has the ability to use descriptive attributes from any domain (i.e., it can natively

Algorithm 2 Finding the best fitting Markov chain order

```
    Input A dataset $G$, a penalty $p$ from $\{AIC, BIC, AICc\}$, start parameter $s$
    Output The estimated Markov chain parameters, the Markov chain order
 1: procedure BestFittingOrder
 2:     ${}_s\mathbf{A}^G \leftarrow$ MarkovChain(G, order = $s$)
 3:     score$_s \leftarrow \phi_{LL}(G, {}_s\mathbf{A}^G, p)$        ▷ Equation (4.7), with penalty term replaced by $p$
 4:     counter = 1
 5:     while $f < s$ do
 6:         $\ell = s - f$
 7:         ${}^{\ell}_{s}\tilde{\mathbf{A}}^G \leftarrow$ normalize(${}_s\mathbf{A}^G$)
 8:         score$_\ell \leftarrow \phi_{LL}(G, {}^{\ell}_{s}\tilde{\mathbf{A}}^G, p)$        ▷ Equation (4.7), with penalty term replaced by $p$
 9:         if score$_\ell <$ score$_s$ then
10:             return ${}^{\ell+1}_{s}\tilde{\mathbf{A}}^G, \ell + 1$
11:         else
12:             counter = counter + 1
13:             score$_s$ = score$_\ell$
14:     return ${}^{\ell}_{s}\tilde{\mathbf{A}}^G, \ell$
```

handle any mix of attributes that are binary, categorical or numerical without the requirement for static pre-algorithm discretization). The algorithm (see Section 2.3.2 and [54, Algorithm 1, page 60]) performs a level-wise search of d levels, where at each level the descriptions of a set of candidate subgroups are further refined and evaluated with a quality measure. The w best-scoring subgroups are selected for the next level. In the end, the algorithm outputs a list of the top-q subgroups.

In order to find subgroups with varying order Markov chains, we have to take a few additional steps to evaluate candidate subgroups. First, we have to find the Markov chain order that best fits the entire dataset Ω. Algorithm 2 describes the procedure. As explained in Section 4.2.2, a higher order Markov chain can be transformed into a lower order model by normalizing the transition matrix. We use this feature to calculate the transition probabilities only once using a Markov chain of order s (line 2), where s is a user-defined parameter which we call the *start* parameter. In line 3, we calculate the penalized log likelihood given penalty $p \in \{AIC, BIC, AICc\}$ as described in Section 4.4. In lines 5-13, we repeatedly normalize the transition matrix (line 7), calculate the new score (line 8) and check whether the score has increased or not (line 9). The procedure returns the parameter estimates and the Markov chain order of the model that maximizes the penalized log likelihood fit.

Algorithm 3 describes how a candidate subgroup is evaluated. First, procedure BestFittingOrder is repeated for the subgroup (line 2). Then, the subgroup is evaluated with quality measure φ_{AIC}, φ_{BIC} or φ_{AICc} (Equation (4.10), (4.11) and (4.12) respectively), depending on parameter $p \in \{AIC, BIC, AICc\}$. Like usual, beam search returns the q best scoring subgroups.

The worst-case computational complexity of the beam search algorithm is $\mathcal{O}(dwZE(c + M(n,m) + \log(wq)))$. Here, d, w, q are as explained earlier, Z is the number of descrip-

Algorithm 3 Evaluating a subgroup with varying order Markov chains

```
Input A subgroup $SG$, a penalty $p$ from $\{AIC, BIC, AICc\}$ and according QM $\varphi$,
    dataset parameters ${}^{\ell}_{s}\tilde{\mathbf{A}}^{\Omega}$, start parameter $s$
Output Real number expressing the exceptionality of subgroup $SG$
procedure EvaluatingSubgroup
    ${}^{u}_{s}\tilde{\mathbf{A}}^{SG}, u \leftarrow$ BestFittingOrder($SG, p, s$)
    quality $\leftarrow \varphi(SG, {}^{u}_{s}\tilde{\mathbf{A}}^{SG}, {}^{\ell}_{s}\tilde{\mathbf{A}}^{\Omega})$
    return quality
```

tors and E the worst-case number of nominal values (numerical and binary descriptors are refined faster). Parameter c refers to the complexity of comparing two models. In our approach, we compare the fit of the subgroup under two models (θ^{SG} and θ^{Ω}) but the parameters of the data model are calculated only once at the beginning of the beam search. The term $M(n,m)$ refers to the cost of learning a model M from n records on m targets. In case of Markov chains, this would compare to fitting a k$^{\text{th}}$ order model for S state-values on N sequences of length T. The computational complexity is then a linear function of N, T and the number of free parameters K, which grows exponentially with base S and exponent s (Section 4.2.2).

The fact that we can evaluate lower order Markov chains by normalizing higher order transition matrices is a powerful feature that keeps the computational complexity of our approach tractable. Still, parameter s is an important parameter because if s is too large, fitting the Markov chain may take unnecessarily long; if s is too small, it is hard to evaluate more complex models. Note furthermore that s determines which parts of the sequences are used for model fitting. After all, fitting a k$^{\text{th}}$ order Markov chain can only be done with the data from time points $k+1$ to T. Although it is possible to normalize higher order transition matrices all the way down to a 1$^{\text{st}}$ order Markov model, the drawback of such a procedure is that not all observations are used for estimating the probabilities.

4.5 Experiments on synthetic data

In the following, we assess internal validity of our proposed method with experiments on synthetic data. For varying data characteristics, we create ground truth subgroups and analyze whether they are ranked first in the top-q result list. Specifically, Section 4.5.1 analyzes exceptional transition behavior, Section 4.5.2 analyzes exceptional starting behavior and Section 4.5.3 contains a sensitivity analysis.

4.5.1 Exceptional transition behavior of varying order

Experimental methodology

We generate synthetic data with $N = 100$ sequences with $T \in \{10, 50, 200\}$ time points, $S \in \{2, 5, 10\}$ states and $Z \in \{5, 10, 20\}$ binary, descriptive attributes.[2] The descriptive attributes are sequence-level attributes, as explained in Section 4.2. For each sequence,

[2] Source code available at `https://github.com/RianneSchouten/simulations_markov_chains_emm/`.

$p(a_z = 1) = 0.5$ for $z \in \{1,2,...,Z\}$. A ground truth subgroup is defined for sequences where $a_1 = 1 \land a_2 = 1$. Thus, approximately 25% of the sequences are part of the true subgroup. All other sequences follow a 1st order Markov chain with probabilities drawn from a uniform probability distribution. The probabilities are normalized to sum to 1. In line with the assumption of stationary sequences, the first time points are sampled using a normalization as well.

Two types of subgroups are generated, as specified by parameter *order* $\in \{1,2,3,4\}$.

1. If order = 1, the subgroup has an exceptional 1st order transition model. This means that both the subgroup and the rest of the dataset follow a 1st order Markov chain, but the parameter values of the subgroup are different. We can write that ${}_1\mathbf{A}^{SG} \neq {}_1\mathbf{A}^{\Omega}$.

2. If order = k for $k \in \{2,3,4\}$, the subgroup follows a kth order Markov model. This means that the subgroup is best modeled with $S^k \times S$ transition matrix ${}_k\mathbf{A}^{SG}$ while the rest of the data is modelled with a 1st order transition model ${}_1\mathbf{A}^{\Omega}$. Here, Algorithm 2 will fit a 1st order Markov chain to the entire dataset, and the subgroups should be fitted with a more complex model.

Every combination of simulation parameters is repeated $nreps = 50$ times.

Given a synthetic dataset with a ground truth subgroup, we perform EMM with 6 different quality measures. First, we apply the three quality measures based on information-theoretic scoring functions as proposed in Section 4.4: φ_{AIC}, φ_{BIC} and φ_{AICc}. We compare our quality measures against three reference measures as mentioned in Section 4.3: ω_{tv}, φ_{WD} and φ_{WRL} (Equation (4.4), (4.5) and (4.6) respectively).

Since ω_{tv} is a parameter-based quality measure, we cannot use it to evaluate subgroups of varying order. Instead, we will first determine the Markov chain order of the entire dataset by applying Algorithm 2 with $p = AIC$ and then evaluate candidate subgroups using that same order. Similarly for φ_{WRL}. In case of φ_{WD}, $p = AIC$ when determining the Markov chain order of the dataset, but candidate subgroups are evaluated with $p = none$. Note that since 75% of the sequences are generated with a 1st order Markov chain, it is unlikely that we will evaluate subgroups against higher order models. However, determining the order of the entire dataset is an important step when analyzing real-world data (see Section 4.6).

For each quality measure, we save the rank of the ground truth subgroup in the $q = 20$ output of the extended beam search algorithm. Since every dataset undoubtedly contains 1 ground truth subgroup, we expect the quality measures to give a first rank to that subgroup. We furthermore check the estimated Markov chain order of the ground truth subgroup and calculate the percentage of simulation repetitions where the correct order is given. If the result list does not contain the ground truth subgroup, we set the rank to $q + 1$ and the order to NaN.

The other parameters of the beam search are $w = 25$ and $d = 3$. We furthermore constrain the subgroup to minimally contain 10% of the sequences. Start parameter $s = 4$.

Results

Figures 4.2 and 4.3 respectively show the rank and the order of the ground truth subgroup. We present the results for $Z = 20$ descriptive attributes. Our information-theoretic based quality measures φ_{BIC}, φ_{AIC}, and φ_{AICc} give similar results and they are therefore presented in one row. Clearly, they give a first rank to the ground truth subgroup when the sequences are relatively long ($T \geq 50$). For shorter sequences ($T = 10$), the subgroup is sometimes, but not always, ranked first. Here, it is easier to find the ground truth subgroup if the state-space is small, the ground truth order is close to 1 and the descriptive space is small (the latter is not visible in Figure 4.2).

Doubtlessly, the ground truth order of the subgroup can only be detected if there are enough observations. First, the estimation of a k^{th} order Markov chain requires $T > k$ time points, and $M > S^k(S-1)$ observations. Second, a larger subgroup allows for a more precise estimation of the Markov chain. For instance, for a state-space of 2, a subgroup with a 1^{st}, 2^{nd}, 3^{rd}, or 4^{th} order Markov chain can be found when the sequences are long ($T = 200$, Figure 4.3). However, when the state-space increases or the number of time points decreases, it may not be possible to detect higher order models (even when the subgroups rank first). Correct estimation of the Markov chain order could possibly also be obtained by increasing the sequence length or the number of sequences N.

Next, quality measures φ_{AIC} and φ_{AICc} are slightly more robust for the number of observations than φ_{BIC}. We can see this in Figure 4.3, where φ_{AIC} finds subgroups with a $k = 2$ Markov chain order when $S = 5$ and $T = 50$ or $S = 10$ and $T = 200$, or with a $k = 3$ order when $S = 5$ and $T = 200$. In contrast, quality measure φ_{BIC} finds the order in none of those subgroups. These findings seem logical since we know that the BIC uses a larger penalty than AIC (Section 4.4). We see no important differences between φ_{AIC} and φ_{AICc}.

We know that φ_{WD} does not use a penalty. Therefore, the estimated order will always be equal to start parameter s, since a more complex model will have a better log likelihood fit. The only limitation is $M < K$, which happens when, for instance, $T = 50$, $S = 10$, and $s \in \{3,4\}$. Consequently, in Figure 4.3, 100% of the subgroups with a true 2^{nd} order Markov model are found, but none of the subgroups with a true Markov model of other orders. Similar results are obtained when $S = 5$ and $T = 200$.

Although φ_{WD} fits the wrong 2^{nd} order Markov chain to all subgroups when $T = 50$ and $S = 10$, almost all subgroups are still ranked first (Figure 4.2). By contrast, when $T = 50$ and $S = 5$, subgroups are also estimated with the wrong model (3^{rd} order), but these do not end up in the top-20 result list. We obtain similar results for $S = 10$ and $T = 200$ (Figure 4.2). Apparently, for $S = 10$, the availability of longer sequences amplifies the difference between the subgroup and the entire dataset, and estimating the wrong order therefore disturbs the ranking, while for $S = 5$, the availability of longer sequences allows for a low ranking even though the estimated Markov chain order is wrong.

The last two rows in Figures 4.2 and 4.3 display results for ω_{tv} and φ_{WRL}. Both measures evaluate subgroups using the order as estimated on the entire dataset. When $T = 200$, the estimated dataset order sometimes equals the subgroup order. Then, ω_{tv} and φ_{WRL} correctly estimate the ground truth Markov chain order (Figure 4.3).

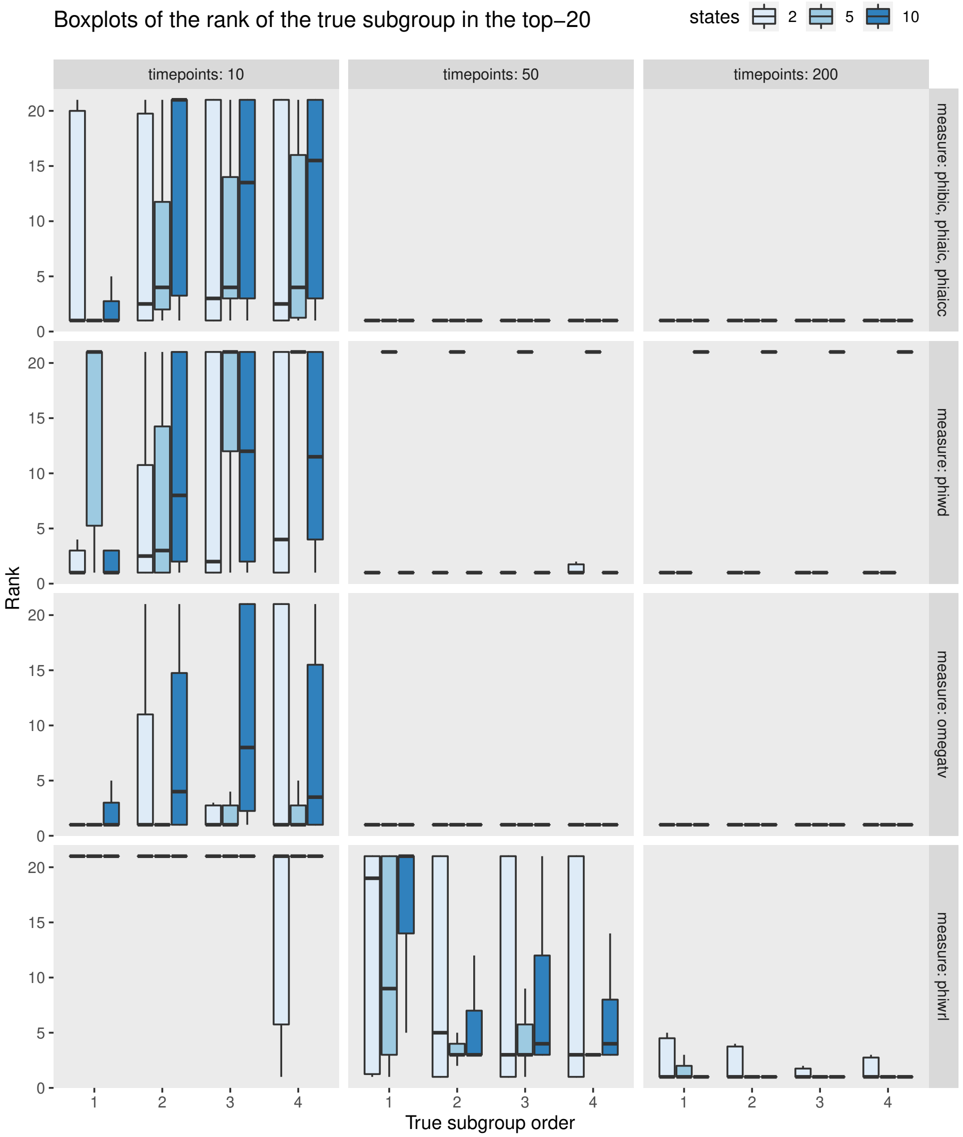


Figure 4.2: Boxplots of the rank of the ground truth subgroup. The ideal value is a rank of 1. We present the simulation results for 20 descriptive attributes $(Z = 20)$ and 50 repetitions $(nreps = 50)$. Results for $\varphi_{BIC}, \varphi_{AIC}$, and φ_{AICc} are similar and therefore presented in one row.

It may still be surprising that ω_{tv} and φ_{WRL} give a first rank to ground truth subgroups with a higher order Markov chain model (Figure 4.2). The likely reason for ω_{tv} is as follows. As discussed in Section 4.4, the difference between a normalized and a directly estimated 1^{st} order transition model is that the latter uses the observations of all available time points

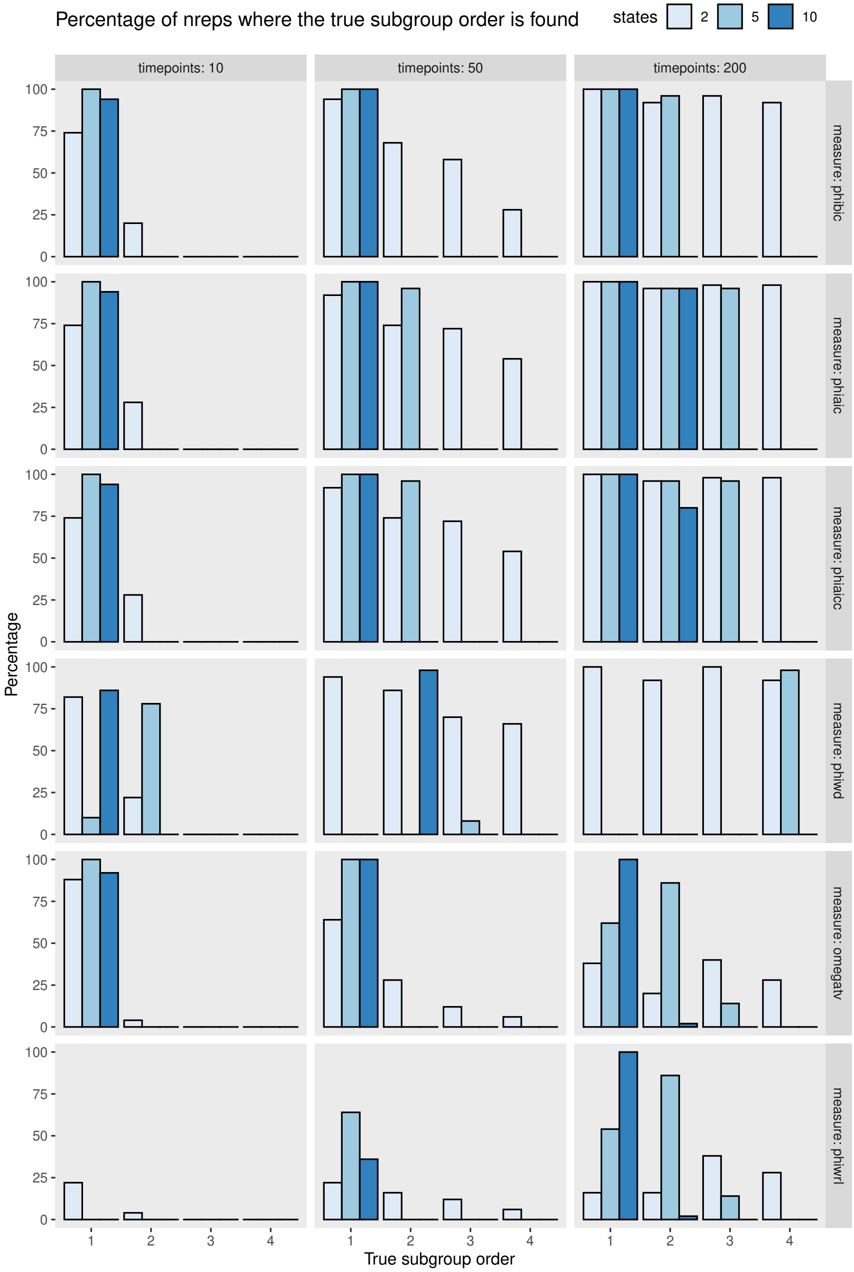


Figure 4.3: Percentage of the number of simulations where the true order of the subgroup is found. The ideal value is 100%. We present the simulation results for 20 descriptive attributes ($Z = 20$) and 50 repetitions ($nreps = 50$).

whereas the first uses only time points $k+1$ to T. For long sequences, this difference is negligible and therefore ω_{tv} (which uses all time points) finds parameter estimates that are close to reality. However, for short sequences with a large number of states, noise disturbs the estimation.

Quality measure φ_{WRL} greatly relies on the parameters that are estimated on the entire dataset. We see that the higher the ground truth order, the more frequently the ground truth subgroup is ranked first ($T = 50$, Figure 4.3). Possibly, when the subgroup follows a higher order Markov model, the dataset parameters are more directed towards the subgroup's complement than when the subgroup follows a 1st order Markov model.

4.5.2 Exceptional starting behavior

Experimental methodology

We further evaluate subgroups of sequences with exceptional *initial* probabilities, or exceptional starting behavior. Specifically, the subgroup follows the same 1st order transition model with the same parameter values as the rest of the data: ${}_1\mathbf{A}^{SG} = {}_1\mathbf{A}^{\Omega}$, but it has a distinct set of initial probabilities (Equation (4.3)), which should not be modeled with normalized probabilities but with a separate set of probability values: $\pi^{SG} \neq {}^0_1\tilde{\mathbf{A}}^{SG}$. We thus reject the assumption of stationary sequences (see Section 4.2.2).

Here, the number of free parameters in the subgroup is $S(S-1)$ for the transition probabilities and an additional $S-1$ free parameters for the initial probabilities vector π. In practice, such a model only makes sense when sequences are very short. We therefore decide to run the simulation for exceptional starting behavior with parameters $N \in \{100, 500, 1000\}$, $S \in \{2, 5, 10\}$, $T \in \{2, 5, 10\}$, $Z \in \{5, 10, 20\}$, $s = 1$ and $nreps = 25$. Search parameters are $q = 20$, $w = 25$, $d = 3$ and subgroups should cover ≥ 10% of all sequences.

Results

Subgroups with exceptional starting behavior follow the same 1st order transition model as the rest of the dataset, but have a distinct pattern for the very first time point. Figures 4.4 and 4.5 present our findings for a state-space of 5. In general, the smaller the state-space, the more advantageous the result.

It turns out that the log likelihood based quality measures (either with or without penalty) perform comparably in ranking the ground truth subgroup. Therefore, these measures are shown in a single row in Figure 4.4. These quality measures give a first rank to the ground truth subgroup when 1) there are enough sequences, 2) the sequences are not too long and 3) there are not too many descriptive attributes. Although it is difficult for evaluation measures φ_{AIC}, φ_{BIC} and φ_{AICc} to give a first rank to a ground truth subgroup with exceptional starting behavior, especially if the sequences are long, these information-theoretic scoring functions do allow for a correct estimation of the Markov chain order (see Figure 4.5). The reason is that exceptional starting behavior causes an increase in the number of free parameters (see Section 4.2.2). As a result, for short sequences ($T = 2$), the penalties are too large to counter-effect the increase in log likelihood. For long sequences

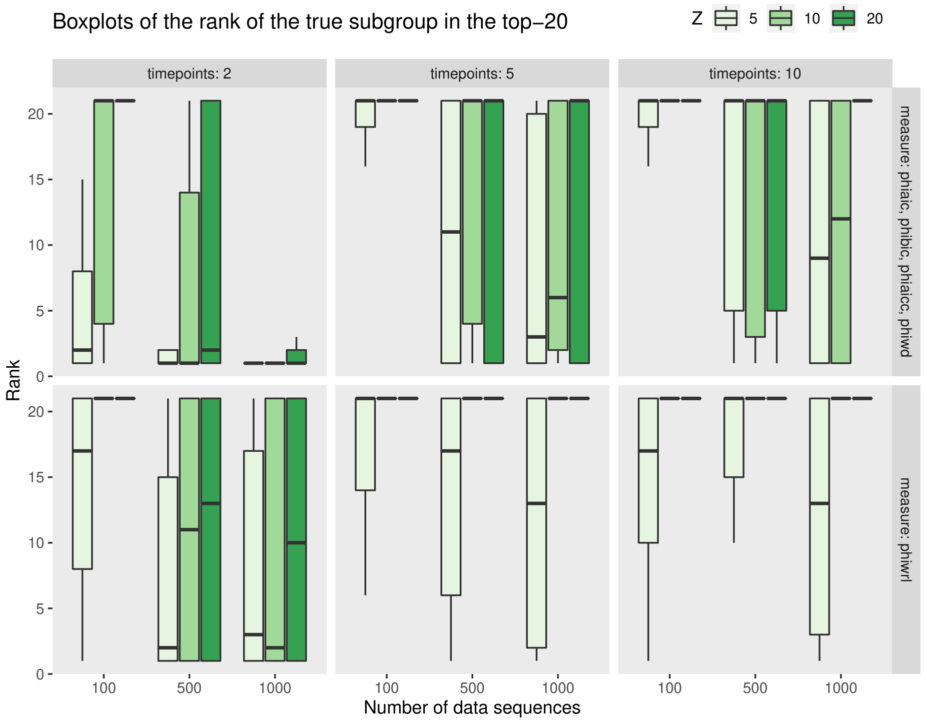


Figure 4.4: Boxplots of the rank of the ground truth subgroup with exceptional starting behavior. The ground truth subgroup differs from the rest of the dataset by its initial probabilities. The transition behavior of subgroup and dataset is the same. The ideal value is a rank of 1. We present the simulation results for 5 states ($S = 5$) and 25 repetitions ($nreps = 25$). Results for $\varphi_{BIC}, \varphi_{AIC}, \varphi_{AICc}$, and φ_{WD} are similar and therefore presented in one row. Quality measure ω_{tv} cannot detect exceptional starting behavior and is therefore not shown.

however, the model fit increases sufficiently. Logically, since φ_{WD} does not use a penalty, it is good at estimating more complex models (Figure 4.5).

Both ω_{tv} and φ_{WRL} evaluate candidate subgroups using the same Markov chain order as estimated on the entire dataset, which is a 1^{st} order chain without additional initial parameters. Using such a model, ω_{tv} never manages to rank the true subgroup first (and we therefore omit the results from the figures), while φ_{WRL} achieves it sometimes when the sequences are as short as possible ($T = 2$) and there are only $Z = 5$ descriptive attributes (Figure 4.4).

4.5.3 Sensitivity analysis

In Section 4.5.1, we analyzed the performance of the quality measures for the setting where the global model is fitted with a 1^{st} order Markov chain, the start parameter $s = 4$, and the subgroups are fitted with a Markov chain order between 1 and 4. The combination

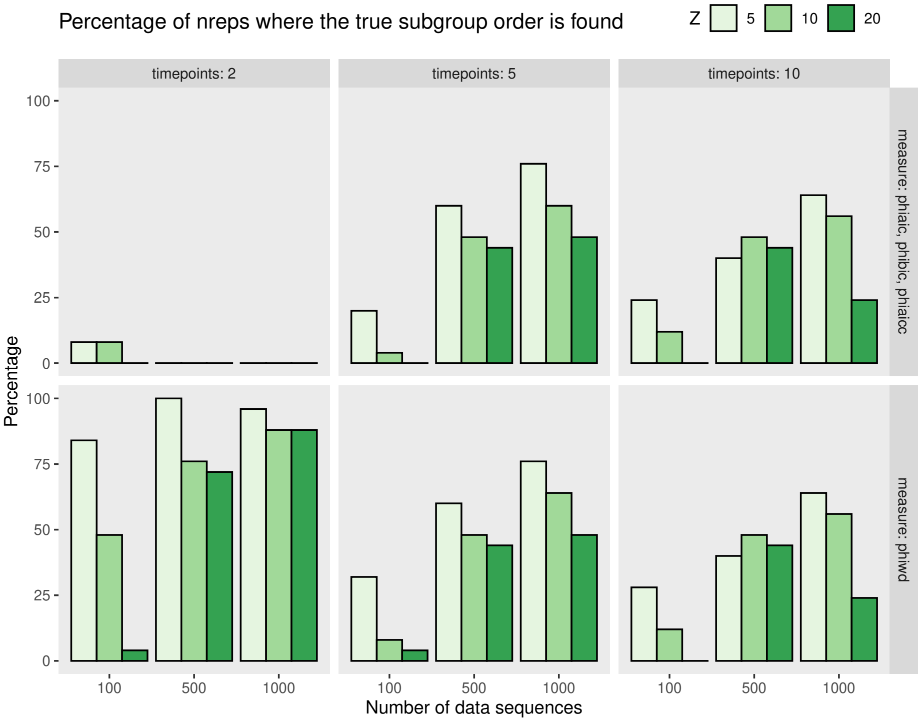


Figure 4.5: Percentage of the number of simulations where the ground truth subgroup is found. The ground truth subgroup differs from the rest of the dataset only by its initial probabilities: transition behavior of subgroup and dataset is the same. The ideal percentage is 100%. We present the simulation results for 5 states ($S = 5$) and 25 repetitions ($nreps = 25$). Results for φ_{BIC},φ_{AIC}, and φ_{AICc} are similar and therefore presented in one row. Quality measures ω_{tv} and φ_{WRL} cannot find the ground truth subgroup order and are therefore not shown.

of these settings allows our algorithms to find the correct order. In this section, we analyze the sensitivity of the results to these parameter settings.

Varying global model order and varying start parameter s

First, we ask ourselves what would happen if the parameters are miss-specified such that the algorithms are steered away from finding the correct order. Specifically, we investigate the effect of:

1. changing the start parameter to $s = 2$,
2. changing the global model to a 3^{rd} order Markov chain.

Therefore, we sample $N = 100$ sequences with a state-space of $S = 5$ and $Z = 20$ descriptive attributes. We vary the length of the sequences with $T \in \{10, 50, 200\}$. The simulation is repeated $nreps = 10$ times. The beam search settings are as before.

Table 4.1: Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{AIC}. The order of the global model is either 1 or 3, and the search is started with parameter $s = 2$ or $s = 4$. We show the results for subgroups where $order \in \{2,3\}$ and for sequences with length $T \in \{10,50,200\}$. Further, $N = 100$, $Z = 20$, $S = 5$, and $nreps = 10$.

		$T = 10$		$T = 50$		$T = 200$	
		True SG order		**True SG order**		**True SG order**	
Gl.order	**Start s**	2	3	2	3	2	3
1	2	2(16)	3(3)	1	1	1	1
	4	11(17)	13(20)	1	1	1	1
3	2	21	21	1	21(11)	1	1
	4	21	21	1	21	1	1

Tables 4.1 and 4.2 show the median and interquartile range (IQR) of the rank of the ground truth subgroup for quality measures φ_{AIC} and φ_{WD} respectively and for subgroups with $order \in \{2,3\}$. We first inspect the results for φ_{AIC} for sequences where $T = 10$. It is clear that when the global model is fitted with a 1st order Markov chain, it is advantageous to set the start parameter to $s = 2$ instead of $s = 4$. In Table 4.1, we see that the median rank decreases from 11 (13) to 2 (3) for subgroups with a 2nd (3rd) order Markov chain. It is surprising that we can give a high rank to 3rd order subgroups using start parameter $s = 2$. Consistent with earlier findings, apparently it can happen that subgroups are considered exceptional even when their Markov chain order is wrongly estimated. Note that these findings hold when T increases.

When the global model is a 3rd order Markov chain and we start evaluating at $s = 2$, the parameter settings forbid the algorithm to correctly estimate the parameters of the global model. However, we see that when T is sufficiently large, φ_{AIC} still ranks the ground truth subgroup first ($T = 200$, Table 4.1). On the other hand, when $T = 50$, it is difficult to find subgroups with a 3rd order Markov chain (but considering the IQR of 11 when $s = 2$, some subgroups can still be found).

For φ_{WD} and a global model order of 1, starting at $s = 2$ instead of $s = 4$ does not decrease the median rank, but it does positively affect the interquartile range ($T = 10$, Table 4.2). When the order of the global model increases from 1 to 3, φ_{WD} allows for discovering ground truth subgroups of order 2 when $s = 2$ and $T = 50$. However, when $T \leq 50$, these subgroups cannot be found using $s = 4$ (and neither can subgroups with $order = 3$). When $T = 200$, all subgroups can be found (just as was the case for φ_{AIC}).

The results for the other quality measures are not shown here but can be accessed in our repository.[3] In sum, quality measures φ_{BIC} and φ_{AICc} perform similarly to φ_{AIC} (Table 4.1), although the IQR of φ_{BIC} is sometimes a bit larger, especially when the global model order is 3 and $T = 200$. For φ_{WRL}, setting $s = 2$ instead of $s = 4$ is advantageous but only when the global model is a 1st order Markov chain. When the order of the global model is 3, φ_{WRL} has trouble finding the ground truth subgroup. Even though ω_{tv}, like φ_{WRL}, greatly depends on the estimated order of the global model, ω_{tv} has large IQRs. This indicates

[3] All results available at `https://github.com/RianneSchouten/simulations_markov_chains_emm/`.

Table 4.2: Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{WD}. The order of the global model is either 1 or 3, and the search is started with parameter $s = 2$ or $s = 4$. We show the results for subgroups where $order \in \{2,3\}$ and for sequences with length $T \in \{10,50,200\}$. Further, $N = 100$, $Z = 20$, $S = 5$, and $nreps = 10$.

		$T = 10$		$T = 50$		$T = 200$	
		True SG order		True SG order		True SG order	
Gl.order	Start s	2	3	2	3	2	3
1	2	4(12)	21(4)	1	1	1	1
	4	5(19)	21	21	21	1	1
3	2	21	21	1	21	1	1
	4	21	21	21	21	1	1

that even when the subgroup order is wrongly estimated, the subgroup can still be found. We saw similar results in Figure 4.2.

Altogether, for shorter sequences, it can be advantageous to decrease the start parameter s. This applies both to 1st and 3rd order global models. In addition, when the global model has a 3rd order Markov chain, the ground truth subgroup can still be found as long as there are enough observations. This holds even when the starting parameter is set to two, which forbids the algorithm from considering the correct order.

Varying subgroup size and varying description length

Second, we investigate the effect of subgroup size and description length on the performance of our quality measures. Therefore, we vary:

1. the description length with $L \in \{1,2\}$,
2. the probability $p(a_z = 1) = pr$ with $pr \in \{0.35, 0.5\}$ for $z \in \{1,2,\dots,Z\}$.

Recall that in Section 4.5.1, $pr = 0.5$ and $L = 2$, resulting in a subgroup that contains 25% of all sequences. Here, we will evaluate subgroups with a coverage of 13% ($pr = 0.35, L = 2$), 25% ($pr = 0.5, L = 2$), 35% ($pr = 0.35, L = 1$) and 50% ($pr = 0.5, L = 1$). We vary the number of descriptive attributes with $Z \in \{5,10,20\}$, set parameters $N = 100$, $S = 5$, and $T = 50$, and we model the global model with a 1st order Markov chain. Like before, we start our search at $s = 4$ and set $q = 20$, $w = 25$, $d = 3$ and the minimum subgroup size to 10%. Consequently, theoretically it should be possible to discover all subgroups. We run the simulation $nreps = 10$ times.

Tables 4.3 and 4.4 present the results for quality measures φ_{AIC} and φ_{WRL} respectively, for subgroups with $order \in \{1,4\}$. We do not show the results for quality measures φ_{BIC} and φ_{AICc}, since they give similar results as in Table 4.1. In Table 4.3, we see that in almost all simulation settings, φ_{AIC} gives a first rank to all subgroups (with an IQR of 0). When $L = 2$ and $pr = 0.5$, it is a bit harder to find the true subgroup, although it is often still ranked second. Clearly, when the true subgroup is large, it is more difficult to distinguish the subgroup from its complement.

Table 4.3: Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{AIC}. The subgroup description has $L \in \{1,2\}$ attributes, where the probability per attribute is $pr \in \{0.35, 0.5\}$. We furthermore vary the number of descriptors $Z \in \{5,10,20\}$ and set $N = 100$, $T = 50$, $S = 5$ and $nreps = 10$. Results are presented for subgroups with a true Markov chain $order \in \{1,4\}$.

		$Z = 5$		$Z = 10$		$Z = 20$	
		True SG order		True SG order		True SG order	
Desc.length	**Prob.**	1	4	1	4	1	4
1	0.5	2(1)	2(1)	1(1)	2	2(1)	1
	0.35	1	1	1	1	1	1(1)
2	0.5	1	1	1	1	1	1
	0.35	1	1	1	1	1	1

For φ_{WRL}, we have already seen that it can find subgroups with a higher order Markov chain (cf. Section 4.5.1). In Table 4.4, we see the same effect. In addition, we see that the larger the subgroup, the easier it is for φ_{WRL} to distinguish the exceptional sequences from the other sequences. For instance, for a subgroup where $order = 4$, and when $Z = 20$, the median rank increases from 1 to 2, to 3, and finally to 18 when the subgroup size decreases from 50% to 35%, to 25%, and to 13%.

Results for φ_{WD} and ω_{tv} can be found in the repository.[3] Essentially, we find that ω_{tv} is fairly robust for subgroup size. For φ_{WD}, we see a pattern: the smaller the Z, the easier it is to find the true subgroup. The subgroup size does not seem to influence the ranking.

In sum, our quality measures based on information-theoretic scoring functions give stable results for ground truth subgroups of varying size. For very large subgroups that contain ≥ 50% of the sequences, the rank increases slightly but not worryingly much.

4.6 Experiments on public, real-world data

In the next section, we apply our proposed methodology to the MovieLens dataset. In Section 5.5, we extensively assess external validity of our proposed methodology.

4.6.1 MovieLens

The MovieLens 100K dataset[4] consists of 943 users, each rating at least 20 movies on an integer scale from 1 to 5. We consider sequences of ratings per user where $20 \leq T \leq 737$ with an average (SD) sequence length of $T = 203$ (139). The Markov chain uses the movie rating values as its state space, so we have $V = \{1,2,3,4,5\}$ and $S = 5$. Specifically, we search for subgroups of users with exceptional rating patterns based on demographic information (age, gender, occupation) and sequence information (sequence length T). The idea behind using sequence length as a descriptive attribute is to form subgroups of users who rate a lot or subgroups of users who rate relatively little. As described before, a user's rating sequence is either entirely part of a subgroup, or not; we do not split sequences.

[4] The MovieLens 100K dataset is available at `https://grouplens.org/datasets/movielens/`.

Table 4.4: Median (interquartile range) of the rank ($q = 20$) of the ground truth subgroup using quality measure φ_{WRL}. The subgroup description has $L \in \{1,2\}$ attributes, where the probability per attribute is $pr \in \{0.35, 0.5\}$. We furthermore vary the number of descriptors $Z \in \{5,10,20\}$ and set $N = 100$, $T = 50$, $S = 5$ and $nreps = 10$. Results are presented for subgroups with a true Markov chain $order \in \{1,4\}$.

		$Z = 5$		$Z = 10$		$Z = 20$	
		True SG order		True SG order		True SG order	
Desc.length	**Prob.**	1	4	1	4	1	4
1	0.5	11(14)	1	12(20)	1	2(20)	1
	0.35	3(11)	1	12(12)	1	18(18)	2(3)
2	0.5	3	3	3(10)	3(1)	3(1)	3(1)
	0.35	15(9)	8(6)	15(9)	10(9)	21	18(15)

The extended beam search algorithms are performed with parameters $q = 20$, $w = 25$, $d = 3$, $b = 4$, and $s = 4$. Subgroups should cover at least 10% of all users. We adopt the following anti-redundancy techniques from [201] as outlined in Section 2.3: a Weighted Coverage Scheme (WCS) with $\gamma = 0.9$, Description-Based Selection (DBS) with a fixed size of $2w = 50$ and Dominance-Based Pruning (DBP). We use quality measure φ_{AIC} to evaluate candidate subgroups.

The entire dataset is best fitted with a 2$^{\text{nd}}$ order Markov chain. In the top-20, we find subgroups of users with either a 1$^{\text{st}}$ or a 3$^{\text{rd}}$ order Markov chain. For instance, the best-scoring subgroup selects users with *occupation* $\neq$ {*other, technician*} $\wedge$ *183* $\leq$ *sequence length* $\leq$ *737* (cov: 16%) and is best fitted with a 3$^{\text{rd}}$ order Markov chain. In subgroup 2, on the other hand, users with short sequences where *20* $\leq$ *sequence length* $\leq$ *73* $\wedge$ *occupation* $\neq$ *technician* are selected (cov: 52%) and a 1$^{\text{st}}$ order Markov chain is fitted.

The results show that a quality measure that takes into account the number of free parameters, such as φ_{AIC}, allows for a flexible evaluation of candidate subgroups. It is reasonable to assume that the entire MovieLens dataset is fitted with a 2$^{\text{nd}}$ order Markov chain as a compromise between shorter and longer sequences. Evaluating candidate subgroups based on such a 2$^{\text{nd}}$ order model (as would be done in the traditional EMM framework using a parameter-based quality measure) would reduce the probability of finding patterns in subgroups that contain short, or long, sequences. Although the MovieLens dataset seemingly does not encompass meaningful relations between user demographics and sequence length, such patterns may exist in other datasets and can be searched for using quality measures based on information-theoretic scoring functions.

4.7 Discussion

We proposed a method for mining sequences with exceptional transition behavior of varying order using quality measures based on information-theoretic scoring functions. On average, the quality measures based on information-theoretic scores outperform the other measures; they give a higher rank to ground truth subgroups, they find the correct Markov chain order more often and they are able to detect subgroups that would otherwise not have been found (Section 4.5.1). In datasets with many, short sequences, exceptional starting behavior can be detected (Section 4.5.2). For long sequences, our quality measures perform robustly with respect to the order of the global model, the start parameter, the subgroup size, and the description length (Section 4.5.3).

In some situations, other quality measures can be valuable as well. For instance, if subgroups are expected to have a similar Markov chain order as the global model, quality measure ω_{tv} performs fine (but the information-theoretic based measures do not perform worse). In situations where the subgroups are expected to have a (much) higher Markov chain order than the global model, semi-evaluation measure φ_{WRL} can be used. Note that φ_{WRL} requires a relatively large number of observations in order to extract the subgroup. The performance of quality measure φ_{WD} is a bit unpredictable, possibly due to its sensitivity to start parameter s.

The quality measures based on information-theoretic scoring functions are flexible and can detect subgroups whose Markov chains 1) have the same order as the global model and 2) have a deviating order. In practice, the global model is an average over all sequences in a dataset, and quality measures that use a penalty based on the number of observations M and the number of free parameters K are able to go beyond such an average. In our study, we have chosen three common penalties; AIC, AIC with small sample correction, and BIC. However, our proposed EMM framework allows for the extension to other penalized scoring functions in a straightforward way.

Interestingly, our findings from controlled experiments do not show much difference between φ_{AIC}, φ_{AICc}, and φ_{BIC}. The first two slightly outperform φ_{BIC} when the number of observations and the number of free parameters is large (Section 4.5.1), due to the excessive penalization by the BIC scoring functions.

It is a bit unexpected that we do not see a difference between φ_{AIC} and its variant for small sample sizes φ_{AICc}. When $1 < M/K < 40$, the penalty in φ_{AICc} is supposed to do more justice to the uncertainty of parameter estimates than the penalty in φ_{AIC} [28, 87, 193]. This means that we would expect φ_{AICc} to give a larger penalty than φ_{AIC}. A possible explanation for the absence of the effect of such a penalty is that as soon as the true subgroup is found, it is so distinctive from the other sequences that a larger penalty does not bother the ranking. Another possible reason could be that with our simulation parameters, we have not been able to capture the dataset characteristics for which such a penalty would make a difference. Nevertheless, in our synthetic data experiments there are many subgroups where $M/K > 40$ and then, the difference between AIC and AICc disappears nonetheless [28, 87, 193].

For all our experiments, we use the extended beam search algorithm as presented in Section 4.4.1. It is generally known that beam search may discover redundant subgroups. Therefore, we applied the following three methods from [201] during our real-world data experiments: Description-Based Selection (DBS), a Weighted Coverage Scheme (WCS) [117] and Dominance-Based Pruning (DBP) (see Section 2.3.3). Note that in the synthetic data experiments, we designed the simulation such that the descriptive attributes do not overlap in coverage (i.e., binary only). Hence, redundancy did not play a role and we were able investigate the ranking of the ground truth subgroup in a more controlled manner.

We performed the description-based selection using a fixed size of $2w$. In our implementation, description-based selection of candidate subgroups occurs before cover-based selection. We think that it is reasonable to assume that starting the latter with $2w$ subgroups allows for a beam that contains w diverse subgroups. We furthermore decided that γ should not be too small in order to not be too rigorous with decreasing the quality of subgroups that have redundant coverage. We therefore set $\gamma = 0.9$.

The beam search algorithm requires a set of parameters that may come across as arbitrary. In general, we suggest to choose the parameter values such that the result list is practical and meaningful. For one thing, this means that the result list should be diverse [201], but it also means that subgroup descriptions should not be too long or a subgroup should not be too small. We have chosen to set $d = 3$ in order to allow for descriptions that contain at most three attributes. These descriptions can easily be remembered and interpreted by the domain expert and are therefore practical. Furthermore, subgroups should be substantially large in order to adopt separate policies or treatment schemes; it seems reasonable to form subgroups that cover at least 10% of the population.

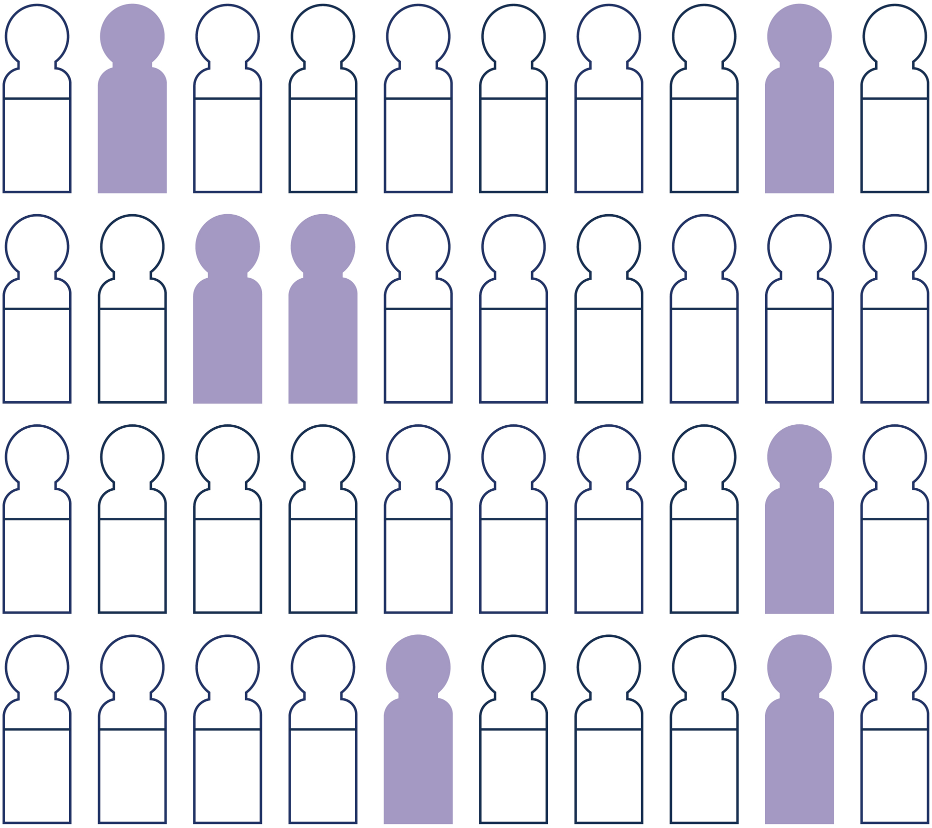
Next, because in both synthetic and real-world data experiments, the number of descriptive attributes is relatively small, we deemed $w = 25$ to allow for sufficient exploration of the search space. For much higher dimensional datasets, possibly this parameter can be increased at some additional computational expense. Last, parameter q is often determined in consultation with domain experts. In our experience, a top-20 result list is not too long to prevent interpretation but long enough to find valuable subgroups. Note that changing q will not actually change the results; it merely specifies the cutoff point in a list of ordered subgroups.

4.8 Conclusion

In sum, we proposed a method for mining sequences with exceptional transition behavior of varying order. Specifically, we use the framework of Exceptional Model Mining (EMM) to find subgroups of sequences and propose a model class for varying order Markov chains. Our model class allows for discovering subgroups in situations where the order of the Markov chain differs between the subgroup and the dataset. Such a situation requires the comparison of a different number of parameters. We therefore do not use a parameter-based quality measure as is common in EMM, but propose three new quality measures based on information-theoretic scoring functions: $\varphi_{AIC}, \varphi_{BIC}$, and φ_{AICc}.

Our findings from controlled experiments show that all three proposed quality measures find exceptional transition behavior of varying order. They all give a first rank to the ground truth subgroup when sequences have a length $T \geq 50$. For shorter sequences, the ability to give a first rank to the ground truth subgroup depends on the state-space, the descriptive space and the ground truth Markov chain order. Naturally, the higher the Markov chain order of the subgroup, the more observations are needed. Nevertheless, φ_{AIC}, φ_{BIC}, and φ_{AICc} all seem sensitive enough to detect the correct Markov chain order but sensitive enough to prevent over fitting. Compared to φ_{BIC}, we find that φ_{AIC} is slightly more robust for the number of observations. We have not seen important differences between φ_{AIC} and φ_{AICc}.

5



Discovering Exceptional Blood Glucose Fluctuations

In this chapter, we discover subgroups of patients with exceptional blood glucose transition behavior. We analyze sequential data from the second DIAbetes and LifEstyle Cohort Twente (DIALECT-2). DIALECT-2 is an observational study of adult persons with diabetes type 2, where blood glucose is measured using the FreeStyle Libre sensor, an intermittently continuous glucose monitoring (iCGM) sensor. In this study, we discretize the continuous measurements into five blood glucose levels and then analyze transition patterns between these levels in two ways: 1) we analyze the full sequences and 2) we calculate the percentage per day that a patient has blood glucose values that are in range, below range, or above range. With both approaches, we discover a variety of subgroups with exceptional Markov chain transition behavior, and a variety of blood glucose fluctuations patterns. Our findings demonstrate the clinical and practical relevance of the approach proposed in Section 4.4 and support clinicians in establishing individualized glycemic treatment.

The contents of this chapter have previously appeared in **Schouten, R. M., Bueno, M. L., Duivesteijn, W., and Pechenizkiy, M.** Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery 36* (2022), 379–413 [170].

5.1 Introduction

The clinical accepted standard for monitoring glycemic control for patients with diabetes type 2 is to measure the blood level of glycated haemoglobin (HbA_{1c}) [208]. However, the use of HbA_{1c} has important limitations. For instance, its assessment does not contribute to reduction of hypoglycemic episodes and it does not reflect blood glucose fluctuations well [38, 108].

An alternative way of assessing blood glucose fluctuations is by taking measurements with the FreeStyle Libre sensor, an intermittently continuous glucose monitoring (iCGM) sensor. It is hypothesized that monitoring blood glucose values using an iCGM device may become the new way to monitor glycemic treatment for patients with diabetes type 2 [38, 46]. Consequently, the following question arises: *How can iCGM-derived parameters support establishing individualized glycemic treatment?*

In this chapter, we use the framework of Exceptional Model Mining (EMM) [54, 121] to further explore the use of iCGM-derived parameters to establish individualized glycemic treatment. In particular, we deploy the method proposed in Section 4.4 ([170]) and aim to discover subgroups of patients with exceptional blood glucose fluctuations.

We do this by discretizing continuous measurements into five blood glucose levels: below range 2 (BR_2), below range 1 (BR_1), in range (IR), above range 1 (AR_1), and above range 2 (AR_2), where range refers to desired blood glucose values. We then analyze transition patterns between these levels in two ways: 1) we analyze the full sequences and 2) we calculate the percentage per day that a patient has blood glucose values that are in range, below range, or above range.

We discover a variety of subgroups with exceptional Markov chain transition behavior, and a variety of of blood glucose fluctuations patterns. That is, some subgroups transition towards higher blood glucose values; others towards similar or lower values. For instance, patients with a high HbA_{1c} value are more likely to have a Time Above Range (TAR) that is too high. If those patients are also older than average, they are additionally less likely to have a good Time In Range (TIR). In contrast, patients with a low HbA_{1c} value are likely to transition away from high blood glucose levels.

Domain experts confirm that our findings are clinically relevant and practically useful; they contribute to further establishing individualized glycemic treatment.

5.2 Background

We assume a dataset Ω with n independently but not identically distributed sequences of discrete random variables $X_t : t \in \{1, 2, ..., T\}$. The data realization at time t is denoted with x_t. Although sequence $r \in \Omega$ has length T_r, without loss of generality, we assume one fixed length T for every sequence. We refer to N as the data *size* and write M to denote the total number of *observations*, where $M = \sum_{r \in \Omega} T_r = NT$ (note that the total number of *transitions* is $M - N$). The set of possible discrete values is $V = \{v_1, v_2, ..., v_S\}$ for all x_t.

We further assume the availability of an extra set of attributes with information about the sequences. We call these attributes descriptors. The full form of sequence r then becomes

$(x_1, x_2, ..., x_{T_r}, a_1, a_2, ..., a_m)$ for all $r \in \Omega$. Here, integer m simply denotes the number of descriptive attributes.

Compared to the generic terminology for EMM (introduced in Section 2.3), here we define a subgroup such that a sequence (e.g., of iCGM measurements) is either covered by a description or not. We do not allow for a sequence to be split into pieces and to be partly assigned to a subgroup. The reason is that we want to find to which patients the exceptional sequences belong; this could assist doctors in adopting appropriate policies. An interpretable description of patients could assist doctors in selecting the most useful treatment; descriptions that only partly include certain patients are less helpful.

5.3 Continuous blood glucose sequences

We analyze data from the second DIAbetes and LifEstyle Cohort Twente (DIALECT-2) [46, 67]. DIALECT-2 is an observational study of adult persons with diabetes type 2, where blood glucose is measured every 15 minutes for a period of 14 days. In general, blood glucose values are considered to be in the desired range (IR) if they are between 3.9 and 10.0 mmol/L. [38] furthermore distinguish blood glucose values that are *below* (BR) and *above* range (AR). These lower and upper ranges are again subdivided into BR_1 (3.0 - 3.9 mmol/L), BR_2 (<3.0 mmol/L), AR_1 (10.0 - 13.9 mmol/L) and AR_2 (>13.9 mmol/L).

The DIALECT-2 dataset contains the information of 126 patients, with an average sequence length of $T = 1210$ (SD: 158). Not all sequences have the same length because sometimes patients forget to upload the stored data or to charge the iCGM-device. On average, 55 (SD: 38) measurements were missing, and no patient had more than 312 missing values.

Overall, the DIALECT-2 dataset has the following characteristics: $N = 126$, $T = 1344$, $M \approx 170000$, and $V = \{\text{BR}_1, \text{BR}_2, \text{IR}, \text{AR}_1, \text{AR}_2\}$ and $S = 5$. As numerical descriptive attributes, we use age, diabetes duration, body mass index, waist/hip ratio, predicted muscle mass, systolic blood pressure, diastolic blood pressure, heart rate, alcohol intake and smoking pack years. The binary descriptors are sex, whether or not someone uses insulin, and if so, with what type of scheme, whether or not someone uses metformin, repaglinide or sulphonylurea, the presence of micro vascular disease and the presence of macro vascular disease. We use one ordinal descriptive attribute: HbA_{1c} category. A HbA_{1c} value ≤ 53 mmol/mol is considered *low*, a value from 54 to 62 mmol/mol *medium* and a value ≥ 63 mmol/mol *high* [11, 134].

5.4 Experimental setup

A quality measure quantifies the difference between behavior within the subgroup and behavior within the entire dataset (or the subgroup's complement) (Definition 2.2). Based on our findings from controlled experiments in Section 4.5, in this chapter, we evaluate subgroups' exceptionality using φ_{AIC}, a quality measure based on Akaike's Information Criterion (AIC) [3, 4].

We apply the (extended) beam search algorithm with parameters $w = 25$, $d = 3$ and $q = 20$. Descriptive attributes are refined with standard strategies (cf. [54]), where we treat numerical attributes with the dynamic discretization strategy lbca from [135] using $b = 4$

bins. Here, lbca is a concatenation of *Local* discretization timing, *Binary* interval type, *Coarse* granularity, and *All* selection method. Furthermore, we set a minimum subgroup size of 10% and use start parameter $s = 4$, that is, we start with evaluating 4^{th} order Markov chains, compare with 3^{rd} order chains, and so on.

It is generally known that the beam search algorithm can discover redundant subgroups. Therefore, we implement three techniques from [201] as outlined in Section 2.3.3. We apply Description-Based Selection (DBS) with a fixed size of $2w = 50$, a Weighted Coverage Scheme (WCS) with $\gamma = 0.9$ and Dominance-Based Pruning (DBP).

5.5 Experimental results

We pre-process the data in two ways, resulting in two result sets. In Section 5.5.1, we first analyze the full (long) sequences of blood glucose values. Next, in Section 5.5.2, we derive the percentage per day that patients have blood glucose values that are above, below or in the desired range. This approach results in short sequences.

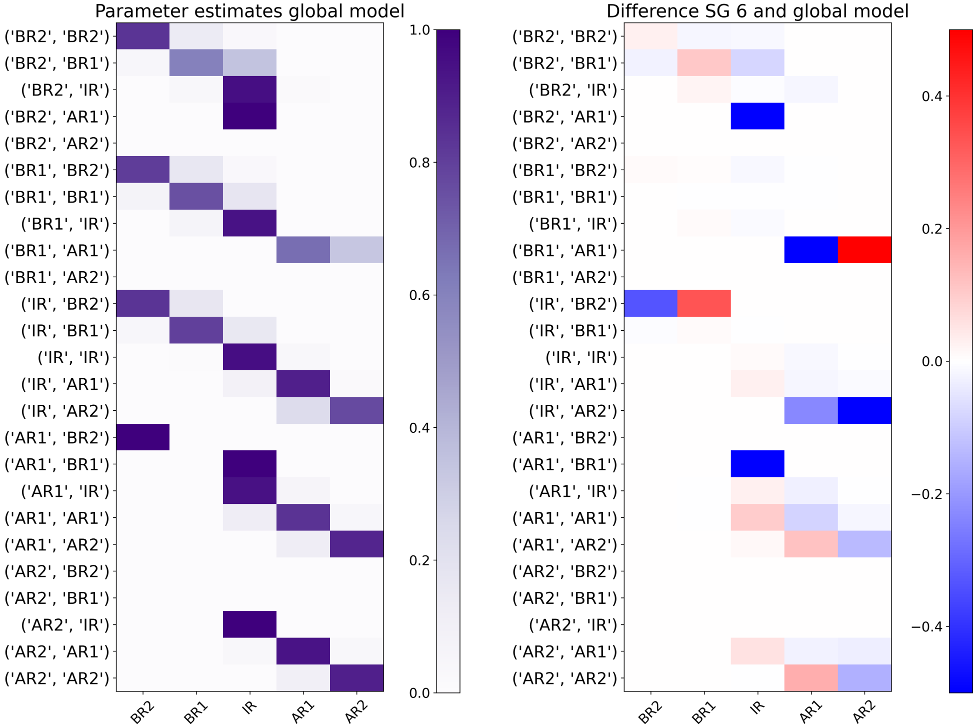


Figure 5.1: Parameter estimates of the global model (left; reproduction of Figure 4.1a for purposes of easy comparison with the figure on the right) and the difference between the sixth best-scoring subgroup and the global model (right). Description is *HbA*$_{1c}$ *category* = *low* ∧ *alcohol intake* ≤ *25 units/month*. Coverage: 24%.

5.5.1 Long sequences of discrete blood glucose level

Our first experiment aims to discover subgroups of patients with exceptional blood glucose fluctuations, as measured by the parameters of a Markov chain fitted to full sequence of discretized blood glucose values (measurements for 14 days with sampling every 15 minutes). Initial results were presented before in Figure 4.1.

In DIALECT-2, the entire dataset is best modeled with a second-order Markov chain; we illustrate this once more in the left subplot of Figure 5.1. Here, we see both diagonal patterns and unusual fluctuations such as blood glucose values changing from IR → AR_2 and from AR_1 → BR_2.

The top-20 subgroups are best fitted with either a 1^{st} or 2^{nd} order Markov chain. The first subgroup contains 18% of the patients with description *HbA_{1c} category = low ∧ diabetes duration ≤ 20 years ∧ 21.3 ≤ BMI ≤ 35.7*. The subgroup's parameter estimates were already shown in Figure 4.1b. Fairly, the conditions for diabetes duration and BMI cover all patients that are in the first three quartiles of the respective variable distributions; they only remove a few extreme patients from the full subgroup. However, the first condition that selects patients with a low HbA_{1c} value is very interesting as we know that HbA_{1c} correlates with the average blood glucose concentration of the past few months and increases the risk for comorbidities [208].

For the first subgroup, we find a strong diagonal transition pattern (i.e., people tend to stay at the same blood glucose level) and the blood glucose values of these patients fluctuate less than those in the overall patient population. This may also be the reason that a 1^{st} order Markov chain suffices. Furthermore, if these patients have a too high blood glucose value (AR_1 or AR_2), there is a chance that they will transition towards the average (i.e. towards AR_1 and IR, respectively, see the fourth and fifth row in Figure 4.1b).

The second-best-scoring subgroup selects patients with a *high* HbA_{1c} value. Figure 5.2 shows the difference in parameter estimates between subgroup 2 and subgroup 1. It is immediately visible that these patients are more likely to transition towards higher blood

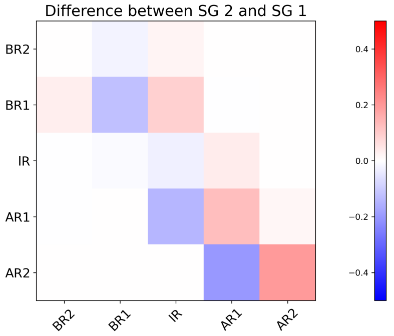


Figure 5.2: Difference between parameter estimates of second best-scoring subgroup and first best-scoring subgroup (Figure 4.1b). Description is *HbA_{1c} category = high ∧ 30.9 ≤ fat percentage ≤ 60.3 ∧ 118.7 ≤ syst.bp ≤ 158.7*. Coverage: 24%.

glucose levels than patients in subgroup 1 (see red squares for $BR_1 \rightarrow IR$ and $IR \rightarrow AR_1$) and less likely to transition to lower levels (see blue squares).

The right plot in Figure 5.1 presents the difference in parameter estimates between the sixth subgroup and the global model, both best fitted with 2nd order Markov chains. Here, like for subgroup 1, patients with *low* HbA_{1c} values are selected. Although we see more fluctuations than for subgroup 1, we see a similar trend where blood glucose levels are likely to stay the same (e.g., the red squares in the first and last two rows), or transition towards a lower blood glucose level (e.g., red for $(AR_1, AR_1) \rightarrow IR$ and $(AR_1, AR_2) \rightarrow AR_1$).

5.5.2 Short sequences of TIR, TBR, and TAR

For our second analysis, we derive the percentage per day that a patient has blood glucose values at level BR_2, BR_1, IR, AR_1, or AR_2. This is referred to as the Time In Range (TIR), Time Below Range (TBR), and Time Above Range (TAR) [11, 38, 46]. For each of these values, we compare the percentages with the guidelines (see Table 5.1) [38]. Subsequently, for each day, we assign one out of 8 state-values (see Table 5.2). This gives us one sequence of length $T = 14$ per patient.

The entire dataset is best modeled with a 1st order Markov chain, because the sequences are relatively short and the dataset size is relatively small. In general, patients stay in or move towards state AC (low TIR, good TBR, high TAR) or state AH (good TIR, good TBR, good TAR) (top left plot in Figure 5.3).

The amount of available data for the subgroups is even smaller than for the entire dataset, and it is therefore not possible to find subgroups with higher order Markov chains. The first subgroup contains 24% of the patients with description *$18.2 \leq$ fat percentage $\leq 42.2 \wedge 34.6$ kg $\leq$ predicted mean mass ≤ 65.5 kg $\wedge$ HbA_{1c} category = high*. These patients are likely to transition to state AA, AC and AG (see top right plot in Figure 5.3, red columns), which corresponds to the situation where TAR is too high.

Table 5.1: Conversion of time spent in glucose level ranges into states suitable for Markov chains. Time In Range (TIR), Time Below Range (TBR) and Time Above Range (TAR) are calculated based on whether glucose values are IR, BR_1, BR_2, AR_1 or AR_2. Medically inspired cut-off percentages are taken from [38].

TIR	IR < 70%	IR ≥ 70%
	low	good
TBR	$BR_1 < 4\%$	$BR_1 \geq 4\%$
$BR_2 < 1\%$	good	high
$BR_2 \geq 1\%$	high	high
TAR	$AR_1 < 25\%$	$AR_1 \geq 25\%$
$AR_2 < 5\%$	good	high
$AR_2 \geq 5\%$	high	high

Table 5.2: Conversion of time spent in glucose level ranges into states suitable for Markov chains. Eight state-values are created based on the combination of the TIR, TBR and TAR (cf. Table 5.1).

State	TIR	TBR	TAR
AA	low	high	high
AB	low	high	good
AC	low	good	high
AD	low	good	good
AE	good	high	high
AF	good	high	good
AG	good	good	high
AH	good	good	good

The third best-scoring subgroups covers patients with, among others, HbA_{1c} = *low* (bottom left plot in Figure 5.3). Here, we see transitions AC → AG (TIR is good instead of low) and AE → AF (TAR is good instead of high).

The fourteenth best-scoring subgroup covers patients with a high HbA_{1c} value, who are additionally older than the average patient. These patients not only have a TAR that is too high, but they are also less likely to have a TIR that is good. We see this in the bottom right plot in Figure 5.3 by the blue columns, and by the two red columns for state AA and state AC. Clinicians and domain experts confirm these findings. It is generally accepted that the blood glucose values of older patients are a bit higher since their risk for comorbidities is lower and their life expectancy shorter.

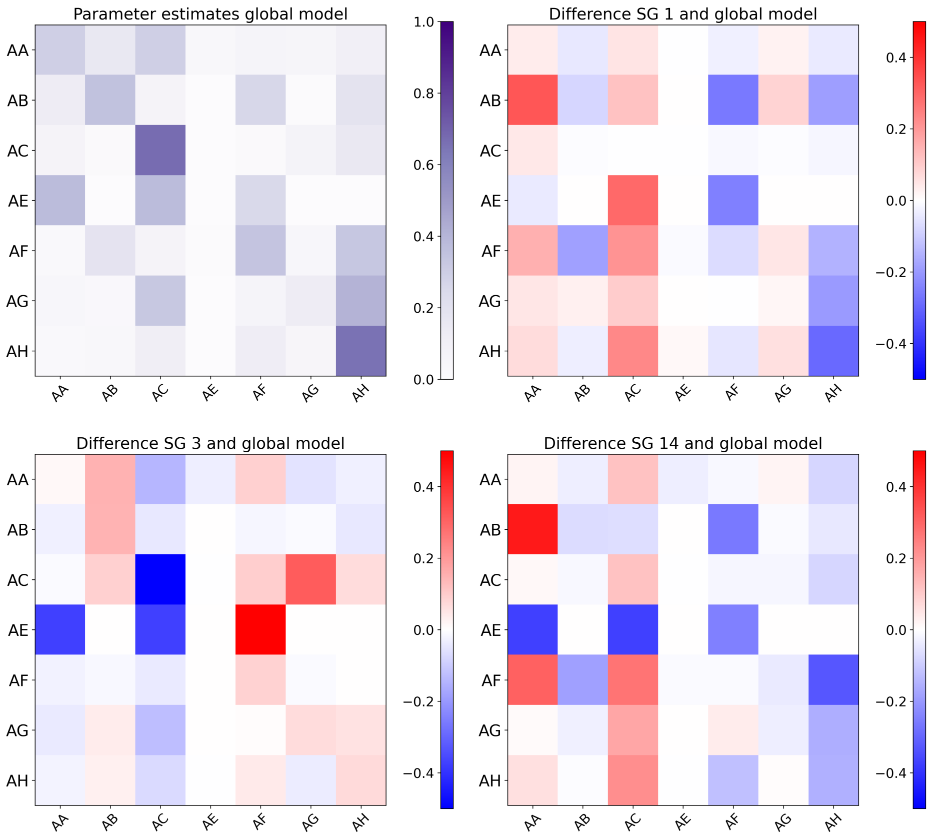


Figure 5.3: Parameter estimates of the global model (top left) and the difference between three subgroups and the global model. Top right: First best-scoring subgroup with description *HbA_{1c} category = high* ∧ *18.2 ≤ fat percentage ≤ 42.2* ∧ *34.6 kg ≤ predicted mean mass ≤ 65.5 kg*. Coverage: 24%. Bottom left: Third best-scoring subgroup with description *HbA_{1c} category = low* ∧ *alcohol intake ≤ 18 units/month*. Coverage: 22%. Bottom right: Fourteenth best-scoring subgroup with description *HbA_{1c} category = high* ∧ *67 ≤ age ≤ 84* ∧ *0.9 ≤ waist/hip ratio* ≤ 1.2. Coverage: 21%.

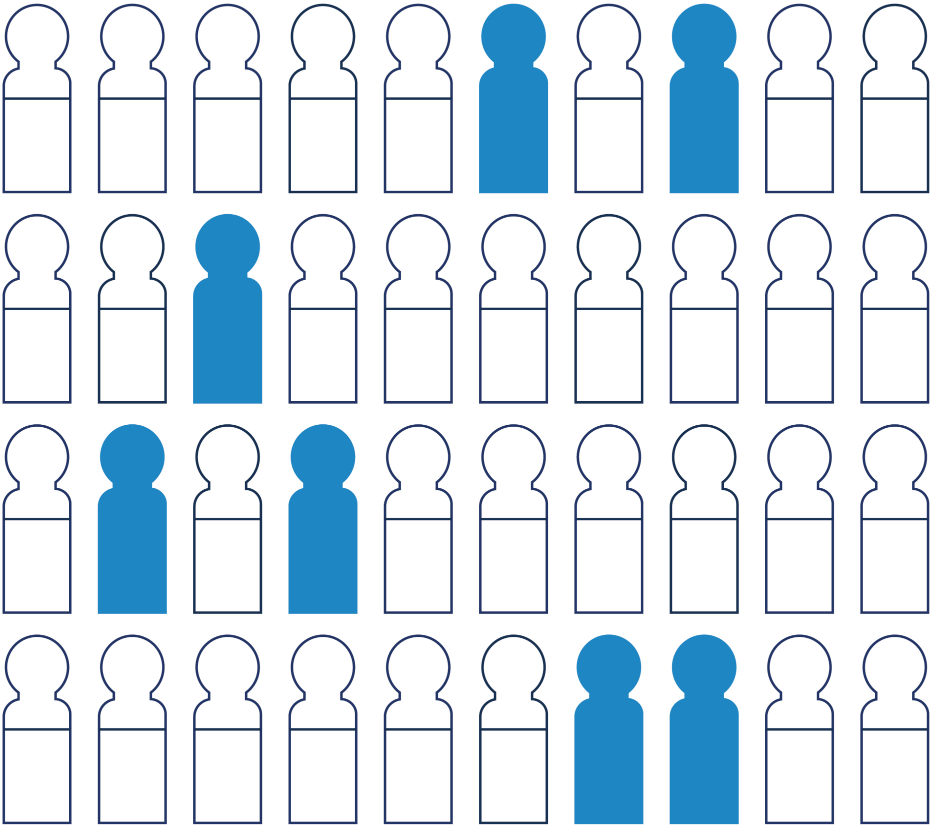
5.6 Conclusion

In this chapter, we demonstrate the practical and medical relevance of the patterns discovered using the approach proposed in Section 4.4. In this chapter, we used real-world data from an observational study of adult persons with diabetes type 2.

In the first experiment, a 2nd order Markov chain is fitted to the entire dataset. We discover subgroups of either a 1st or 2nd order Markov chain. For instance, we find a first subgroup that covers patients with low HbA_{1c} values, a measure known to correlate with average blood glucose values. The subgroup is best modeled with a 1st order Markov chain and its parameter estimates show an increased probability of staying in or moving towards desired blood glucose values. Clinicians and domain experts confirmed that the blood glucose values of these type of patients fluctuate less.

In the second experiment, we find, among others, subgroups covering patients with high HbA_{1c} values and an above average age. The model parameters indicate an increased probability of transitioning to blood glucose values that are too high. Clinicians and domain experts confirmed these findings, and furthermore add that it is generally accepted that the blood glucose values of older patients are a bit higher since their risk for comorbidities is lower and their life-expectancy shorter.

6



Exceptional Model Mining for Repeated Cross-Sectional Data

Repeated Cross-Sectional (RCS) data measures a phenomenon by repeatedly sampling new individuals from a population at successive measurement moments. It allows for analyzing societal trends without the need to follow individuals. To gain a deeper understanding of these trends, we propose EMM-RCS, an Exceptional Model Mining instance designed to find subgroups displaying exceptional trend behavior in RCS data. We build quality measures on the standard error, finding various types of exceptionalities within trends (exceptional flattening, slope, deviation from the norm). Additionally, EMM-RCS can handle practical RCS data problems, including uneven spacing of measurements over time, fluctuating sample sizes, and missing data. We explore the performance of two refinement strategies for incomplete descriptors and demonstrate the performance of our quality measure using synthetic data experiments and two public datasets. Our findings demonstrate the versatility of our generic quality measure.

The contents of this chapter have previously appeared in **Schouten, R. M., Duivesteijn, W., and Pechenizkiy, M.** Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In *Proc. SDM* (2022), pp. 585–593 [171]. We also incorporated most information covered by its **supplementary material** [172].

6.1 Introduction

A deeper understanding of societal trends helps policy makers, government institutions and decision makers to take the right course of action. For instance, consider the trend in the percentage of Dutch adolescents that consumed alcohol in the last 4 weeks, which has decreased from 57% in 2003 to 26% in 2015 and has flattened since then [161, 192]. Since adolescent alcohol consumption has short-term risks (e.g., injuries, violence) and long-term risk of adult alcohol dependence [40], the Dutch government formulated a new goal that by 2040, the percentage of Dutch adolescents that consumed alcohol in the last 4 weeks must have further decreased to 15% [199]. Policy makers are now developing campaigns specifically targeted at the right group of adolescents.

For developing such strategies, it is valuable to gain a deeper understanding of the interplay between various socio-demographic factors such as gender, school level, family situation, and ethnicity [40, 99]. In particular, we want to know the trends in alcohol use in certain subgroups of the population, and where and when those trends are deviating from the general, population trend. In this chapter, we develop EMM-RCS: a new instance of the framework of Exceptional Model Mining (EMM) [54, 121] that seeks subgroups with exceptional societal trends in Repeated Cross-Sectional data.

An introduction to EMM was given in Section 2.3: EMM is the data mining method that is tailored best towards the task of analyzing societal trends: on the one hand, the evaluation within its search strategy allows to find a variety of trend deviations; on the other hand, exploring the search space of subgroups that can be concisely described ensures that the results are interpretable for domain experts, making the translation of data mining results to policy decisions relatively straightforward.

A trend analysis is done by collecting data with a Repeated Cross-Sectional (RCS) research design, also called a trend design [25]. RCS data is obtained by sampling new individuals from a population at successive occasions. It differs from time series where multiple measurements are taken per individual with very short time interval. It also differs from longitudinal data where the same people are followed through time (see Figure 6.1). In-

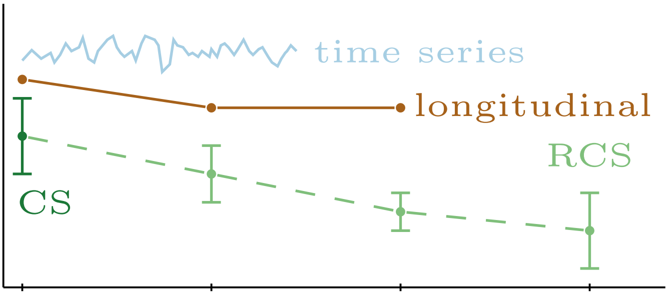


Figure 6.1: Schematic overview of various types of data. In time series, multiple measurements are sampled per individual with very short time intervals. In longitudinal data, sampling is done with long intervals in a relatively long period of time. Repeated cross-sectional (RCS) data is collected from new individuals at each measurement occasion, resulting in varying sample sizes.

stead, RCS research collects the same information from different individuals and therefore allows for the analysis of change over time without the need to follow people. This can be useful in case of dropout risk, or when following participants is not possible (e.g., adolescents grow older).

We cannot directly apply existing instances of EMM to RCS data for a few reasons. Foremost, no model class and quality measure exist that are suitable for analyzing trends. There is some work on EMM for sequential data [131, 138] but as in time series or longitudinal data, there the sequence is known per individual and the sample size is fixed. Instead, in RCS data, an individual contributes to the trend at just one measurement occasion. Consequently, the entire trend is estimated on data with varying sample sizes; an EMM model class and quality measure would have to be able to handle such fluctuations.

In addition, RCS research often has a long-term focus where the interest is in estimating trends for years or decades. Thus, the distribution of descriptive attributes is likely to change over time as well. For instance, the proportion of Dutch adolescents joining secondary school at a high level has increased [30]. On the one hand, EMM should allow for the forming of subgroups even if there is a strong imbalance in the distribution of descriptors, but on the other hand, EMM should also account for the resulting trend estimate uncertainty.

We propose a generic, flexible quality measure that uses the standard error of the trend estimate to account for both fluctuating sample sizes, varying descriptor distributions and uncertainty of trend estimates. By using the standard error we additionally direct the search away from small subgroups. Our quality measure can be used for any trend estimate for which a standard error exists (or can be calculated using bootstrapping).

Moreover, the generality of our quality measure allows to define multiple types of trend deviations as exceptional behavior. This is important from a domain perspective. For instance, when analyzing alcohol usage trends, domain experts are interested in finding subgroups of adolescents who drink more, who have a stronger or weaker decrease, and who have many flat parts in the trend. These different types of deviation may provide different kinds of information, such as to whom the campaign should be targeted, how to design the campaign, and who is likely not to be influenced.

In sum, our main contributions are:

1. an EMM model class for RCS data, including a way to handle missing data in descriptive space and irregular measurement occasions in target space,
2. a generic quality measure that can be adapted for finding various exceptionalities in trends,
3. the use of standard error to handle fluctuating sample sizes, varying descriptor distributions, and uncertainty of trend estimates, while concurrently directing the search away from small subgroups.

6.2 Related work

A Repeated Cross-Sectional (RCS) research design is used in many studies, such as the European School Survey Project on Alcohol and Other Drugs [137], British Social Attitudes (cf. `https://bsa.natcen.ac.uk/`), and Monitoring the Future [94]. In the respective domains, the interplay between socio-demographic factors is investigated using global analysis techniques. For instance, regarding alcohol use among Dutch adolescents, several, separate logistic regression models are employed to test for significant interaction effects between survey year as dummy variable and each of the socio-demographic factors [40]. Such global analysis methods do not allow to explore more than a few socio-demographic factors or to find non-linear effects. Also, variables have to be categorized beforehand and individuals are nested into distinct groups.

Instead, we use the framework of Exceptional Model Mining (EMM) [54, 121] to search for subgroups with exceptional trends. EMM poses no restrictions on the number of descriptive attributes and the type of interaction between those attributes. To the best of our knowledge, we are the first to analyze RCS data using local pattern mining. The vast majority of data mining (and hence also EMM) methods are developed for observational data that is available but not specifically collected with a certain research design, maybe except for a few directions such as uplift modeling that uses experimental data [163] and an EMM model class for A/B tests [52].

Remark that EMM model classes exist for sequential data [131, 170]. Similarly, methods exist to detect time series anomalies or discords [126]. However, in both sequential and time series data the repeated measurements are taken within individuals (cf. Figure 6.1), which requires different methods than analyzing change over time in RCS data (where individuals only contribute to the trend at one measurement occasion).

We propose a generic quality measure that builds on the standard error of the trend estimate. The standard error has been proposed before in an interestingness measure for subgroup discovery (SD) [80] with a numerical target [122], where it is called a *t-score* since it evaluates the mean estimate of a target variable. We use the concept of standard error more flexibly by calculating a *z-score* of any user-defined trend estimate and using it to evaluate an entire trend instead of just one estimate or target attribute. The reader should not confuse our notion of a z-score with what [150] propose as a variant of the t-score and call z-score; they combine the standard deviation of the target attribute in the *entire* dataset with the size of the subgroup, this is not the same as standard error.

6.3 Preliminaries

Repeated Cross-Sectional (RCS) data originate from a quantitative research design where measurements are taken at several occasions, each from a new sample of individuals [25]. One can see an RCS dataset Ψ as a bag of datasets Ω_{x_t} where each dataset is collected at measurement occasion $x_t \in \mathcal{T} = \{x_1, \ldots, x_t, \ldots, x_T\}$. For instance, the Health Behaviour in School-aged Children study (HBSC) [192] collects data with 4-year intervals. Later in this dissertation, we use HBSC data from 2005 to 2017; hence, $\mathcal{T} = \{05, 09, 13, 17\}$. In RCS data, the time interval between x_t and x_{t+1} can be both regular and irregular.

Our goal is to analyze the change over time of a population parameter μ. Conform statistical theory, for a random variable (RV) Y, each sampled value y^i in the dataset represents one of the values $Y^1, Y^2, ..., Y^N$ in the population. An estimator uses the sampled values to estimate the parameter. For instance, $\overline{Y}$ can be used as a point estimator of the mean of a population and $\overline{y}$ is its point estimate [22].

An estimator performs well if it produces unbiased and precise estimates and its performance depends largely on the sampling design [22]. In this chapter, we only use unbiased estimators. The variance of an estimator is an indicator of the amount of variation in the possible outcomes of the estimator. We will use estimators that estimate this variance using the sampled values. Regarding the sampling design, we will assume that every individual i has the same probability of being included in the sample with inclusion probability $\pi^i = {}^{n_{x_t}}/_{N_{x_t}}$ where n_{x_t} is the sample size and N_{x_t} the population size at occasion $x_t \in \mathcal{T}$. Our method can be extended to other sampling designs.

Exceptional Model Mining (EMM) [54, 121] seeks subgroups in a dataset that somehow behave exceptionally. In the context of societal trends, we aim to utilize the EMM framework to discover subgroups in society with exceptionally deviating trends. EMM generally deploys a rule-based description language using conjunctions of attribute-value conditions; by applying this technique to socio-demographic information, EMM could be used to form interpretable subgroups such as *11 ≤ age ≤ 15 ∧ school year = 4 ∧ lives with both parents = yes.*

Notation and definitions for EMM were introduced in Section 2.3. However, the traditional EMM terminology is not directly applicable to RCS data. After all, in RCS data, individuals are nested in measurement occasions and we need to take into account that the observations reside at a lower hierarchical level than the target model. The next section explains our proposed solution.

6.4 Our proposed approach: EMM-RCS

First, we define our notion of data as follows:

Definition 6.1 (RCS data)**.** *An RCS dataset* $\Psi = (\Omega_{x_1}, \ldots, \Omega_{x_t}, \ldots, \Omega_{x_T})$ *is an ordered bag of* T *datasets, where each* Ω_{x_t} *is collected at measurement occasion* x_t *for* $x_t \in \mathcal{T}$*. Every* Ω_{x_t} *is a bag of records* $r_{x_t} \in \Omega_{x_t}$ *of the form* $r_{x_t} = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m)$*. The dataset size is* $n^\Psi = \sum_{t=1}^{T} n_{x_t}$.

The main difference between the general framework of EMM (Section 2.3) and EMM-RCS is the simple addition of a time indicator x_t for record r and dataset Ω. However, the important consequence is that record $r^i_{x_t}$ is only measured at occasion x_t and its values are not known for other measurement occasions. Consequently, the sample sizes differ per occasion; $n_{x_t} \neq n_{x_{t'}}$ $(t \neq t')$. In the rest of this chapter, where possible, we use the word *individual* instead of *record*.

Definition 6.1 assumes that attribute a_j exists for all $x_t \in \mathcal{T}$. In practice, not all RVs will be sampled at every occasion. We use k and m to denote the number of unique descriptive and target attributes in the entire RCS dataset Ψ; any attribute a_j may be absent at any occasion $x_t \in \mathcal{T}$.

6.4.1 Descriptive space

A subgroup is defined as the bag of records that a description covers. A description D covers an individual $r^i_{x_t}$ if and only if $D(a^i_1, a^i_2, \ldots, a^i_k) = 1$ (cf. Definition 2.1). Specifically in RCS data, a description D collects individuals from all measurement occasions, which allows to estimate the trend in the subgroup. Henceforth, we distinguish the entire dataset from a subgroup by superscripts Ψ and SG. Then, the subgroup size is n^{SG} and its coverage ${}^{n^{SG}}/{}_{n^{\Psi}}$.

A complication is that the distribution of attribute a_j may vary over time. For instance, in the Netherlands, the number of adolescents with a non-native background fluctuates [30]. A condition on *ethnic group* may therefore result in a very small sample for a particular measurement occasion. Furthermore, value a_j may not be available for individual $r^i_{x_t}$; whether or not individual $r^i_{x_t}$ is covered by a description is undefined. Values may be missing because attributes were removed from or added to the data collection, or because an attribute may not be applicable to certain respondents. The former type results in missing values for all $i \in n_{x_t}$ individuals $r^i_{x_t}$ at occasion x_t. The second type makes that value a_j could be missing for a specific individual $r^i_{x_t}$, but be observed for another individual $r^{i'}_{x_t}$ at the same occasion x_t $(i \neq i')$.

We decide for two things. On the one hand, we allow for subgroup SG to have a different number of observed occasions than the entire dataset, up to a user-defined minimum constraint c_{occ}. We then say that ${}^{T^{SG}}/{}_{T^{\Psi}} \geq c_{\text{occ}}$, which allows to form subgroups on descriptive attributes that are sampled at many but not all occasions.

On the other hand, we define a new refinement condition for incomplete attributes. The canonical EMM description language uses conjunctions of conditions on single attributes. During the search, a refinement operator η builds a new set of descriptions by looping over all descriptive attributes and adding conditions to existing descriptions (cf. [54, Section 4.1]). We add a condition where the attribute-value pair is missing.

Definition 6.2 (Refining incomplete attribute)**.** *For an incomplete descriptive attribute a_j, construct a response indicator $R_{a_j} \in \{0,1\}$ with $R_{a_j} = 1$ if a value is observed and $R_{a_j} = 0$ if a value is missing. Then add $D \cap (R_{a_j} = 0)$ to the set of descriptions $\eta(D)$.*

In Section 6.7.3, we conduct two experiments to evaluate various applications of this refinement strategy.

For complete descriptors, refinement strategies exist for binary, numerical, and nominal descriptive attributes [54]. In RCS data, distinguishing nominal from ordinal attributes is practically relevant. Hence, we define a refinement strategy for ordinal attributes.

Definition 6.3 (Refining ordinal attribute)**.** *For an ordinal attribute a_j, order the unique values of a_j; this gives a list of ordered values $w_1, \ldots, w_m$. Then, add $\{D \cap (a_j \leq w_h), D \cap (a_j > w_h)\}_{h=1}^{m-1}$ to the set of descriptions $\eta(D)$.*

6.4.2 Quality measure

We analyze trends as a model class and aim to find subgroups with exceptional deviations in that trend. For each description D in description language $\mathscr{D}$, a quality measure quantifies the exceptionality of the trend in the subgroup covered by that description. The top-q EMM task is to find the q best-scoring subgroups for that quality measure.

We propose the following quality measure for finding subgroups with exceptional trends:

$$\varphi_{RCS}(D) = f\left(\left\{z_{x_t} \mid x_t \in \mathscr{T}\right\}\right) \tag{6.1}$$

$$z_{x_t} = \frac{\left|\theta^{SG}_{x_t} - \theta^{0}_{x_t}\right|}{se\left(\theta^{SG}_{x_t}\right)}. \tag{6.2}$$

Our quality measure φ_{RCS} consist of an inner part that measures exceptionality per occasion, and an outer part that summarizes the T values into one overall quality value. Hence, in Equation (6.1) we have $f : \mathbb{R}^{1\times T} \mapsto \mathbb{R}^{1\times 1}$; examples are the maximum, average, or sum. We discuss choices for f and their implications in Section 6.5.2 and now focus on Equation (6.2).

In Equation (6.2), $\theta^{SG}_{x_t}$ is the value of a statistic calculated in the subgroup, $se(\theta^{SG}_{x_t})$ is its standard error, and $\theta^{0}_{x_t}$ is a reference value. The reader may recognize this as a z-score or standard score, and we indeed intend to measure the number of standard deviations that $\theta^{SG}_{x_t}$ deviates from the reference value $\theta^{0}_{x_t}$. Here, we have the flexibility to decide whether we want to find subgroups whose trends deviate from the global trend in the entire dataset, from the trend in the complement of the subgroup, or from a fixed value such as 0.

Also, we can choose a statistic for θ_{x_t}. For instance, to directly evaluate the trend values, we can set $\theta_{x_t} = \mu_{x_t}$, where μ_{x_t} can be any population parameter (e.g., mean, prevalence, ratio). Value μ_{x_t} can thus be estimated using one or more RVs. Instead of directly comparing the trend values, we could also assess exceptional increases or decreases in a trend, or find subgroups for which the trend is stable (cf. Section 6.5.1).

We incorporate the sample size of the data at occasion x_t by setting the denominator as the standard error of the value estimated in the subgroup. The standard error depends on the sampling distribution of the estimator and on the sample size. The larger the sample size, the smaller the standard error and hence the larger the standard score z_{x_t}. Standard error correction will direct the search process away from tiny subgroups. In case the distance to reference value θ^0 is similar at two occasions, but the sample size at x_t is larger than at $x_{t'}$ $(t \neq t')$, more weight will be on the distance at occasion x_t (since we are more certain about that distance). Hence, the search can use descriptive attributes whose distributions change over time, but corrects for imbalance over time by giving more weight to estimates calculated from more data.

6.4.3 Search strategy

We choose to employ *beam search* [54, Algorithm 1] as the algorithm of our choice. Essentially, beam search performs a level-wise search of d levels: on each level, promising descriptions are further refined. Candidate subgroups are evaluated using φ_{RCS} and w promising subgroups are taken to the next level for further refinement. The top-q subgroups are returned (see 2.3.2 for more on beam search).

EMM-RCS requires the computation of T estimates and T standard errors which is unlikely to be more complex than $\mathcal{O}(Tn)$ for most statistical quantities (n being the number of individuals). The computational complexity of the entire beam search algorithm then becomes $\mathcal{O}(dwZE(c+\mathcal{O}(Tn)+\log(wq))$, where we replaced the cost of learning a model M from n records on m targets, $M(n,m)$, with $\mathcal{O}(Tn)$, and where Z and E refer to the number of descriptors and the highest cardinality (number of distinct values) of any nominal attribute, respectively (numerical, ordinal and binary descriptors are refined faster than nominal descriptors).

Furthermore, EMM often employs a constraint that specifies the minimum size of a subgroup. We adapt this constraint such that it checks the sample size at occasion x_t for all $x_t \in \mathcal{T}$, which we denote with c_{size}. This minimum sample size constraint c_{size} is generally set to $c_{\text{size}} = 0.1$. Consequently, if the subgroup has at least a c_{occ} proportion of occasions (as discussed in Section 6.4.1) with at least a c_{size} proportion of individuals, the subgroup will be evaluated. The subgroup's data from occasions that do not meet the size constraint are removed and not used while evaluating the subgroup's exceptionality.

6.5 Instantiations of our quality measure

Before we apply our method to both synthetic and real-world data in Sections 6.6 and 6.7, we will now first give examples of choices for μ, θ, and f.

6.5.1 Instantiations of μ and θ

The Dutch government aims to decrease the proportion of Dutch adolescents that consumed alcohol in the past month [199]. To seek exceptional trends in alcohol use in Section 7.5, we assume a binary-valued RV L measuring alcohol use at occasion x_t, following a binomial distribution with parameters n_{x_t} and μ_{x_t}, assuming that n_{x_t} is large [22]. Then, μ_{x_t} can be approximated with the proportion of the sampled values ℓ, and corresponding standard error,

$$\mu_{x_t} = \frac{1}{n_{x_t}} \sum_{i=1}^{n_{x_t}} \ell^i_{x_t} \tag{6.3}$$

$$se\left(\mu_{x_t}\right) = \sqrt{\frac{\mu_{x_t}\left(1-\mu_{x_t}\right)}{n_{x_t}-1}}. \tag{6.4}$$

While analyzing the Eurobarometer dataset (Section 6.7.2), we are interested in the European citizens' perception about the speed with which the European unification advances. There, we assume that RV L measures the speed on a scale between 1 and 7 and that it has a normal distribution with mean μ. We set μ_{x_t} as in Equation (6.3) with standard error:

$$se\left(\mu_{x_t}\right) = \frac{sd\left(\mu_{x_t}\right)}{\sqrt{n_{x_t}}} = \left(\frac{\sum_{i=1}^{n_{x_t}}\left(\ell^i_{x_t}-\mu_{x_t}\right)^2}{\sqrt{n_{x_t}}\left(n_{x_t}-1\right)}\right). \tag{6.5}$$

If we are interested in finding trends with an exceptional increase or decrease at some measurement occasions, we can decide to set θ_{x_t} as the difference, or slope, between the estimates of two successive occasions. Of course, this would only work if the data for successive occasions exist. For the estimate of Equation (6.3) the slope and its standard error are

$$\theta_{x_t} = \mu_{x_{t+1}} - \mu_{x_t} \qquad \forall t \in \{1,2,\ldots,T-1\} \tag{6.6}$$

$$se\left(\theta_{x_t}\right) = \sqrt{se\left(\mu_{x_{t+1}}\right)^2 + se\left(\mu_{x_t}\right)^2}. \tag{6.7}$$

Sometimes, a trend may fluctuate a little between successive measurement occasions, while the human eye can distinguish a clear general pattern. Then, directly comparing the slope in the subgroup with the slope in the entire dataset may result in finding false subgroups that are considered exceptional because of sampling fluctuations. Hence, one may want to first calculate a weighted moving average τ_{x_t} with a window u,

$$\tau_{x_t} = \frac{\sum_{t=1}^{u} w^*_{x_t}\mu_{x_t}}{\sum_{t=1}^{u} w^*_{x_t}} = \sum_{t=1}^{u} w_{x_t}\mu_{x_t}, \tag{6.8}$$

where $w_{x_t} = {w^*_{x_t}}/{\sum_{t=1}^{u} w^*_{x_t}}$. If we weight our moving average by the respective sample sizes and choose a window of $u = 2$. Then, $w_{x_t} = {n_{x_t}}/{(n_{x_t}+n_{x_{t+1}})}$ for all $t \in \{1,\ldots,T-1\}$. The standard error of this statistic is:

$$se\left(\tau_{x_t}\right) = \sqrt{\sum_{t=1}^{u} w^2_{x_t} se\left(\mu_{x_t}\right)^2}. \tag{6.9}$$

As a second step, we can then define θ_{x_t} and its standard error as in Equations (6.6) and (6.7), with μ_{x_t} replaced by τ_{x_t}. While calculating the weighted moving average with a window of 2, we lose one measurement occasion; another one is lost while calculating the slope. Hence, the number of values entered into Equation (6.1) is $T-2$.

6

6.5.2 Instantiations of f

The function f aggregates T standardized values into one subgroup quality value. When choosing the right function f, we must keep in mind the ordering of the subgroups in the top-q search. Subgroups with a larger quality value are ranked higher, and we consider z_{x_t} to be larger for more exceptional subgroups. Hence, the maximum, average, or sum are appropriate choices for f, but the minimum is not.

The maximum is simply $f_{\max} = \max_{z_{x_t}} z_{x_t}$ for all $x_t \in \mathscr{T}$. If we set the reference $\theta^0_{x_t} = \theta^\Psi_{x_t}$ as the general trend in the dataset, $f_{\max}$ selects subgroups that deviate at least once from the general trend. Instead, one could also take the average over the T standardized scores; $f_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} z_{x_t}$. As can be seen in Section 6.7.2, such a summary function selects exceptional subgroups with smooth trends while $f_{\max}$ results in fluctuating trends. Of course, f_{sum} prefers subgroups for which more measurement occasions are available.

As we will see in Section 7.5, the general trend in alcohol use is predominantly decreasing. We could be interested in finding subgroups of adolescents whose alcohol usage trends have horizontal parts: for those adolescents, government campaigns may fall flat. We can find such subgroups by setting $\theta^0_{x_t} = 0$. However, without any further adaptation, due to the ordering of subgroups in the top-q search, these settings will result in subgroups with slopes that deviate from 0, instead of being close to it. Reversing the ordering won't help, since this directs the search towards smaller subgroups: z_{x_t} in Equation (6.2) decreases if $se(\theta^{SG}_{x_t})$ increases.

We experiment with two solutions. First, we do not correct for varying sample sizes by setting $se(\theta^{SG}_{x_t}) = 1$ and let $f_{\text{count}}(\epsilon) = |z_{x_t} < \epsilon|$ count the number of scores within a threshold ϵ. The higher the count, the more exceptional the subgroup. Second, we do estimate the standard error of the slope, but instead of dividing by the standard error, we multiply the distance by the standard error. Again, we use $f_{\text{count}}(\epsilon)$, although it requires a bit more time to specify the right parameter for ϵ. In combination with such a multiplication, one could also use f_{sum}, f_{avg}, or f_{min} and reverse the ordering in the top-q search. However, this would select subgroups with a trend that is in its entirety close to 0 (for f_{sum} and f_{avg}) or subgroups with a trend that has just a single slope that is close to 0 (f_{min}). Based on these preliminaries experiments, our final solution includes setting $se(\theta^{SG}_{x_t}) = 1$, then an $f_{\text{count}}(\epsilon)$ and then an f_{sum}; we thus combine two aggregation functions.

6.6 Experiments on synthetic data

To show the performance of quality measure φ_{RCS}, we perform a synthetic data experiment as follows. First, we draw trend values from a normal distribution $\mathcal{N}(10,1)$ for $N = 10\,000$ individuals and randomly assign individuals to one out of $T = 10$ measurement occasions. Second, we draw $ncovs = 10$ binary descriptors, each from a binomial distribution $a_j \sim \text{Bin}\left(n = N, p = 0.5\right)$. Third, we generate a ground truth subgroup with a description based on $nlits \in \{2,3,4\}$ literals, which are randomly chosen from the 10 binary descriptors (e.g., $a_4 = 1 \wedge a_7 = 1$).

Because of the way the descriptors are generated, a description with 2, 3 or 4 literals will approximately cover 25%, 12.5% and 6.25% of the individuals. For these individuals, the trend value will be replaced by a new trend value, which is drawn from a normal distribution $\mathcal{N}\left(10 + dist, sd^2\right)$ where the distance varies with $dist \in \{1,2,3\}$ and the standard deviation varies with $sd \in \{1,2,3\}$. The idea is that the standard deviation influences the standard error of the trend estimate. Altogether, these simulation parameters allow us to analyze how the quality value is influenced by varying distance, uncertainty of the trend estimate and size of the subgroup.

Specifically, we perform EMM-RCS with φ_{RCS} with Equations (6.3) and (6.5), $\theta^0 = \theta^{\Psi}$ and f_{max}. We search the space of candidate subgroups using beam search with parameters $q = 20$, $d = 5$, and $w = 20$. We apply Description-Based Selection (DBS), a Weighted Coverage Scheme (WCS) and Dominance-Based Pruning (DBP) (see Section 2.3.3). Every combination of simulation conditions is repeated $nreps = 100$ times.[1]

Figure 6.2 shows boxplots of the quality values of the ground truth subgroups that can be found in the top-20 results list. The smaller the subgroup, the larger the quality value

[1] Experimental code available at `https://github.com/RianneSchouten/EMM_RCS/`.

(compare the dark boxplots for $nlits = 2$ with the lighter boxplots for $nlits = 3$ and 4). Furthermore, the smaller the uncertainty of the trend estimate, the larger the quality value (compare the green boxplots for $sd = 1$ with the orange and purple boxplots for $sd = 2$ and 3). Finally, the larger the distance between the subgroup and global trends, the larger the quality value (compare the two panels).

Figure 6.2 furthermore displays the trade off between the distance, the uncertainty of the trend estimate and the subgroup size in determining the exceptionality of a subgroup. Recall that the larger the quality value, the more exceptional the subgroup. Subgroups with a trend line at $dist = 1$ with $sd = 1$ and $nlits = 2$ (dark green boxplot at the top) have the same quality value as subgroups with $dist = 3$, $sd = 3$ and $nlits = 2$ (dark purple boxplot in bottom panel, see vertical line). Indeed, although the latter are further away from the global trend, the uncertainty is larger. Therefore, their exceptionality cannot be distinguished from subgroups with a trend that is closer but have smaller uncertainty.

Table 6.1 presents additional results. Out of 100 repetitions per combination of simulation conditions, it presents the fraction of ground truth subgroups that is found in the top-20 result list, and the median rank and quality value of the found subgroups. We see that the fraction of found subgroups drops below 1.0 if subgroups are fairly small ($nlits = 4$). Given the way the descriptive attributes are generated, a ground truth subgroup with 4 literals will contain about 6.25% of the individuals. In some repetitions, that percentage may have dropped below the minimum size threshold $c_{size} = 0.05$ and the ground truth subgroup cannot be found.

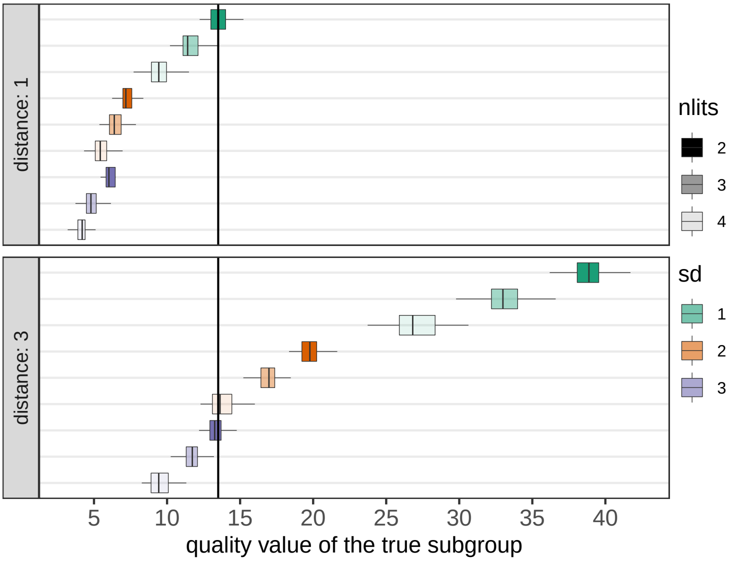


Figure 6.2: Boxplots of the quality values of the ground truth subgroup for 100 repetitions. Top and bottom panel mark the distance between the subgroup and the non-subgroup, $sd \in \{1,2,3\}$ specifies the standard deviation of the trend in the subgroup and $nlits \in \{2,3,4\}$ is the number of literals in the description, which directly influences the size of the ground truth subgroup (25%, 12.5% and 6.25%).

At the same time, it is not always possible to find a larger subgroup. Consider the fraction of found subgroups of 0.06 for $dist = 1$, $sd = 3$ and $nlits = 2$. Here, we remind the reader that φ_{RCS} compares the subgroup trend estimate θ^{SG} with the global trend estimate $\theta^0 = \theta^\Psi$. The latter is an average over both subgroup and non-subgroup individuals in the dataset. The observed distance between θ^{SG} and θ^Ψ will therefore be smaller than the $dist$ intended, depending on the size of the subgroup and the standard deviation in the subgroup. For instance, if the trend values of the subgroup and non-subgroup individuals are both drawn from the same normal distribution and if $dist = 1$, for a subgroup with 25%, 12.5% or 6.25% coverage, the observed distance between θ^{SG} and θ^Ψ will be 0.75, 0.875 and 0.9375 respectively. This effect increases if the standard deviation is larger in either the subgroup or non-subgroup. Therefore, under some simulation conditions, it can be challenging to discover a larger subgroup.

Table 6.1: Fraction of ground truth subgroups that are found in top-20 result list, and the median rank and median quality value of the found subgroups. The subgroup trend has a distance from the non-subgroup trend with $dist \in \{1,2,3\}$, a standard deviation $sd \in \{1,2,3\}$ and a ground truth description of $nlits \in \{2,3,4\}$ literals. The $nlits$ directly influences the subgroup size (25%, 12.5% and 6.25%). Simulation is run with 100 repetitions.

dist	sd	nlits	frac.found	med.rank	med.quality
1	1	2	1.0	1	13.5
		3	1.0	1	11.4
		4	0.56	1	9.4
	2	2	1.0	2	7.1
		3	1.0	1	6.4
		4	0.67	1	5.4
	3	2	0.07	10	6.0
		3	0.92	5.5	4.8
		4	0.47	3	4.2
2	1	1	1.0	1	26.0
		3	1.0	1	22.1
		4	0.52	1	17.6
	2	2	1.0	1	13.4
		3	1.0	1	11.5
		4	0.64	1	9.4
	3	2	0.81	5	9.4
		3	1	1	8.1
		4	0.56	1	6.8
3	1	2	1.0	1	38.9
		3	1.0	1	33.0
		4	0.66	1	26.8
	2	2	1.0	2	19.8
		3	1.0	1	17.0
		4	0.58	1	13.6
	3	2	0.99	4	13.3
		3	0.94	1	11.7
		4	0.63	1	9.4

6.7 Experiments on public, real-world data

In this section, we evaluate our quality measure φ_{RCS} on two real-world datasets, using various combinations of θ and f. We furthermore investigate the effect of new refinement operators for incomplete descriptors in Section 6.7.3.

6.7.1 The Brexit dataset

A 10-wave survey examines Attitudes Towards Brexit (ATB) in the aftermath of the 2016 Brexit referendum on European Union (EU) membership [82]. The survey was conducted between April 25, 2017 and January 10, 2020. The goal of the ATB survey is to examine social identities that are formed during the referendum. Combining the ATB survey with panel datasets, we know that Brexit identities are prevalent, felt to be personally important and cut across traditional party lines [82].

Here, we construct a trend of the proportion of respondents that identify themselves as *leaver* (as opposite to *remainer* or *neither a leaver nor a remainer*). We drop 1 descriptive attribute because it misses ≥ 50% of values. From the resulting 15 descriptors, 6 contain missing values, 1 is binary, 2 are numerical, 6 nominal, and 6 ordinal. The dataset contains 16 965 individuals.

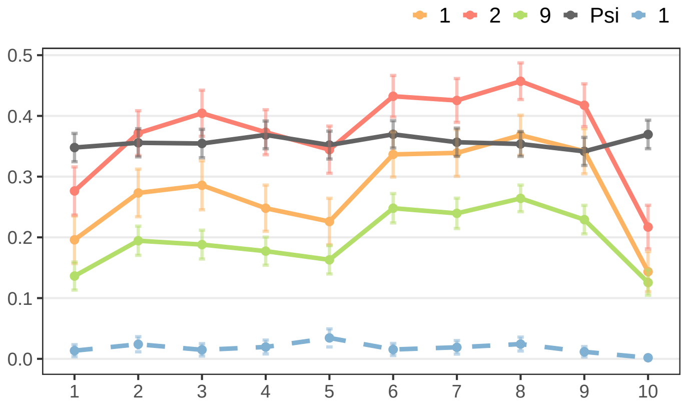


Figure 6.3: Trends of a selection of discovered subgroups in the Brexit dataset, and the overall population (black). Subgroup descriptions are given in Table 6.2.

Table 6.2: Coverage and description of a selection of discovered subgroups in the Brexit dataset. Corresponding trends are displayed in Figure 6.3.

#	Cov.	Description
1	0.32	govthand = {don't know, very, fairly badly} ∧ tradeimmig ≤ 7 ∧ age ≥ 47
2	0.40	govthand = {don't know, very, fairly badly} ∧ age ≥ 39 ∧ work status ≠ {other}
9	0.65	govthand = {don't know, very, fairly badly} ∧ tradeimmig ≤ 7
1	0.34	hindsight = {wrong} ∧ region ≠ East ∧ age ≥ 31

In the Brexit dataset, we explore trends of the proportion of people who think of themselves as *leavers*. Results can be found in Figure 6.3. The population trend is fairly horizontal, with an approximate average of 35% of respondents who want to leave the EU.

The dashed line is the best-scoring subgroup when we directly compare the proportion in the subgroup with the population trend. Dashed subgroup 1 covers people who think in hindsight that Britain was wrong to vote to leave the EU (cf. Table 6.2). The subgroups with the solid trend lines are found by comparing the slopes in the subgroup with the slopes in the population. Now, we find subgroups with an increase in the proportion of leavers at measurement occasions 2 and 3 (first bump) and at occasions 6, 7, and 8 (second bump), but an enormous decrease between occasions 9 and 10. The final occasion was measured on January 10, 2020; a month earlier, on December 12, 2019, the UK General Election delivered a landslide majority for Boris Johnson's conservatives.

Solid subgroups 1, 2, and 9, while sharing fluctuations, appear at different intercepts. The definitions in Table 6.2 show that all subgroups think that Britain is bad at negotiating its future relationship with the EU (condition 1). The other conditions select different age groups (#1, #2) or believe that Britain should prioritize free trade rather than controlling immigration (#9). In general, while the overall population reacted to the 2019 election with a slightly boosted *leave* proportion, in all the subgroups 1, 2, and 9 the *leave* proportion plummeted dramatically. It is quite likely that Boris Johnson's cavalier approach towards all things Brexit and all matters of negotiation has a strongly polarizing effect: those people who already thought that the British government were doing a less than ideal job in negotiations are likely to no longer identify with his particular brand of *leave* politics, while the overall population may be more likely to do so.

6.7.2 The Eurobarometer

The Eurobarometer is a survey organized by the European Commission (EC). Since 1974, the Eurobarometer collects data from citizens in all European Union (EU) countries at 2 or 3 moments per year. We use the public dataset available at [167], containing data from all measurement occasions between 1974 and 2002.

Specifically, we analyze the trend in the mean perception towards advancement of European unification on a scale from 1 (standing still) to 7 (as fast as possible) The survey question is asked at 12 measurement occasions with irregular time intervals: $\mathscr{T} = \{86, 87, 90, 92, 93, 94, 95, 96, 97, 99, 00, 01\}$. This is not problematic since EMM-RCS is built to handle irregular time intervals.

Due to the large scale of the survey, there are no descriptive attributes with no missing values. We drop all descriptive attributes with ≥ 50% missing values, because it is likely that those attributes are surveyed in other years or that they result from follow-up questions (which are only asked depending on a respondent's answer to another question). We apply Definition 6.2 to the resulting 38 descriptors. Of those descriptors, 17 are binary, 1 numerical, 4 nominal, and 16 ordinal. The dataset contains 155 244 individuals.

Figures 6.4 and 6.5 display subgroups with exceptional trends obtained with summary functions f_{avg} and f_{max}, respectively. Tables 6.3 and 6.4 give the subgroup descriptions.

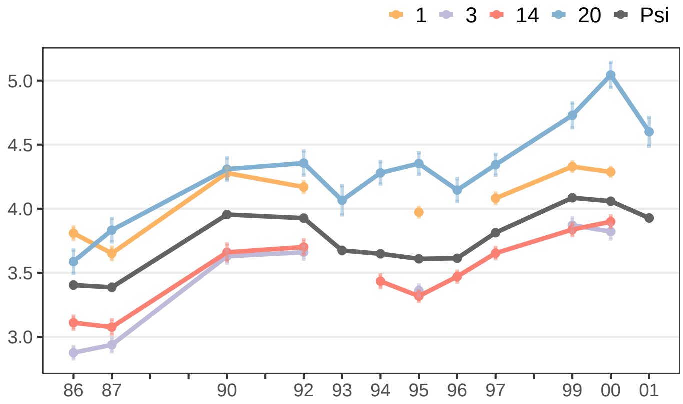


Figure 6.4: Trends of a selection of discovered subgroups in the Eurobarometer dataset using f_{avg}, and the overall population (black). Subgroup descriptions are given in Table 6.3.

Table 6.3: Coverage and description of a selection of discovered subgroups in the Eurobarometer dataset using using f_{avg}. Corresponding trends are displayed in Figure 6.4.

#	Cov.	Description
1	0.17	benefitc = {benefit} ∧ epimp1 = {fairly, very important} ∧ satisdmo = {fairly, very satisfied}
3	0.11	age ≥ 23 ∧ epimp = {not at all, not very important} ∧ satisdmo = {not very, fairly, very satisfied}
14	0.23	benefit = {not benefit}
20	0.06	nation = {Ireland} ∧ married ≠ {separated}

Comparing the two approaches, averaging tends to find subgroups that have a smoother trend, while maximizing tends to find more erratic trends (except maybe for subgroup 20 in Figure 6.4). Most of the subgroups found with f_{avg} deviate from the general trend by translation: fluctuations follow the overall trend, but there is a constant positive or negative intercept based on general group outlook. For instance, subgroup 1 covers European citizens who think their country benefits from being a member of the EU, think the European Parliament (EP) is important in their life and are satisfied with the way democracy works in their country. In contrast, subgroup 3 covers citizens who do not think the EP is important for their life and who are not satisfied with the way democracy works in their country. The former group of citizens is more positive about the advancement of European unification than the latter, which stands to reason.

These findings are important from a domain perspective: they display an interplay between socio-demographic factors that is difficult to find with confirmatory, global analysis techniques. From a DM perspective, these findings illustrate that EMM-RCS detects trends deviating from the global trend in both upwards and downwards directions.

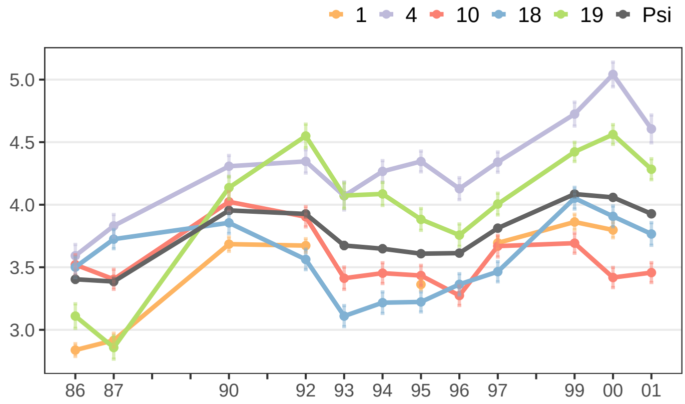


Figure 6.5: Trends of a selection of discovered subgroups in the Eurobarometer dataset using f_{max}, and the overall population (black). Subgroup descriptions are given in Table 6.4.

Table 6.4: Coverage and description of a selection of discovered subgroups in the Eurobarometer dataset using f_{max}. Corresponding trends are displayed in Figure 6.5.

#	Cov.	Description
1	0.12	epimp = {not at all, not very important} ∧ married ≠ {refused} ∧ poldisc = {occasionally, frequently}
4	0.06	nation = {Ireland}
10	0.07	nation = {Italy}
18	0.07	nation = {France}
19	0.07	nation = {Denmark}

Compared to Figure 6.4, the subgroups in Figure 6.5 deviate from the general trend at some points (e.g., #18 in 1993 and #19 in 1992) but overlap with the general trend at other occasions. We find subgroups that barely exceed the minimum size constraint and found subgroups cover the citizens of specific countries. For instance, in Denmark (#19), citizens were more positive between 1992 and 1996 than the average European citizen while in France (#18), the perception of citizens dropped quickly between 1990 and 1992 and then gradually increased again.

Subgroup 20 in Figure 6.4 and subgroup 4 in Figure 6.5 have similar trends and cover mainly the same respondents; citizens in Ireland. We may expect this trend to have a higher rank considering that it is higher on the Y-axis than a subgroup like #1. However, subgroup 1 is larger than subgroup 20 (cf. Figure 6.3) and therefore, its standard error is smaller and its z-scores are larger. Thus, here we see the effect of our correction for imprecise estimates in small subgroups.

The Eurobarometer dataset has missing values for many descriptive attributes and therefore, a subgroup may not have any observed values at some measurement occasions. The effect is visible in Figure 6.4 by interrupted trend lines.

6.7.3 Handling incomplete descriptors

In RCS data, descriptive attribute a_j could be missing for a specific individual $r^i_{x_t}$, but observed for another individual $r^h_{x_t}$ at the same occasion x_t $(h \neq i)$. Popular missing data methods such as dropping incomplete individuals and mean/median imputation are insufficient for solving this issue, for several reasons. The former may simply drop the entire dataset. In less extreme examples, both methods suffer under all forms of missingness: for missingness other than Missing Completely At Random (MCAR), they are known to give biased estimates; under MCAR, they give unrealistic standard errors [128, 196], especially when the proportion of missing values is high (which is typical for RCS data). Moreover, it is hard to impute attribute a_j at occasion x_t if it is not sampled at all at that same occasion.

We therefore introduced an extra refinement operator in Definition 6.2. If the missing data are MCAR, we expect the condition $R_{a_j} = 0$ to not appear in the descriptions of the top-q result list. After all, in case of MCAR every individual has the same, fixed probability of being incomplete and consequently, the distribution of the missing values is similar to the distribution of the observed values. If the data are Missing At Random (MAR), information about missing values lies in the observed data (for other attributes). We expect those other attributes to appear in the descriptions. When the data are Missing Not At Random (MNAR), the information about the missing values is missing from the data. Here, we expect Definition 6.2 to be helpful. For more on MCAR, MAR, and MNAR, see [162, 179, 196].

As we will see in Section 6.7, it turns out that none of the subgroups found in the real-world data experiments selects missing values from a descriptive attribute as proposed in Definition 6.2. To show that these findings are expected if the data is Missing Completely At Random (MCAR) or Missing At Random (MAR), we now present the results of two experiments. In the first experiment, we explore the performance of Definition 6.2; in the second experiment we explore the performance of a *conjunction* of selection conditions that also includes Definition 6.2.

Missingness in the Brexit dataset

First, we experiment with the Brexit dataset. We artificially remove values from the complete dataset, specifically from variable *hindsight* (using techniques from [176]). Recall that *hindsight* is an important attribute for describing subgroups. Table 6.5 presents the main condition for the top-1 subgroup after applying various missingness mechanisms and percentages. Note that about 50% of the individuals in the Brexit dataset have *hindsight* = *wrong*, and that a missingness percentage of 25% therefore means that we have an observed value *wrong* for 25% of the individuals, a missing value for another 25% of the individuals and an observed value *right* for 50% of the individuals.

When data is MCAR, individuals are randomly amputed. Consequently, the distribution of *hindsight* will not change; only the number of observed values will reduce. We see in Table 6.5 that when 15% of the rows have a missing value, not enough observations remain to reliably estimate the trend of the proportion of people who want to leave the EU. Instead, attribute *EURef16*, which has a correlation with *hindsight* of about -0.2, will be used. With a small missingness percentage of 5%, the search is not affected by the missing values.

Table 6.5: Main condition in the description of the top-1 subgroup found on the Brexit dataset; *hindsight* is artificially amputed from a given percentage of rows, for MCAR, MAR, and MNAR missingness mechanisms.

%	MCAR	MAR	MNAR
0.05	h.sight = wr.	h.sight = wr.	h.sight = wr.
0.15	EURef = rem.	h.sight = wr.	EURef = rem.
0.25	EURef = rem.	EURef = rem.	EURef = rem.
0.35	EURef = rem.	EURef = rem.	h.sight = missing
0.45	EURef = rem.	EURef = rem.	h.sight = missing

Table 6.6: Overview of four approaches for handling incomplete descriptors in EMM. Definition 6.2 is used in both the *ignoreORselect* and *ignoreANDselect* method, but in various ways. We give an example of the selection conditions added to the set of descriptions for an incomplete, binary descriptor a_j with values v_1 and v_2.

Method	Description	Added selection conditions
cca	incomplete individuals are removed from dataset	$a_j = v_1$, $a_j = v_2$
ignore	incomplete individuals are ignored and cannot be covered	$a_j = v_1$, $a_j = v_2$
ignoreORselect	incomplete individuals could form a separate subgroup	$a_j = v_1$, $a_j = v_2$, $R_{a_j} = 0$
ignoreANDselect	incomplete individuals could form a separate subgroup, or may be selected together with complete individuals	$a_j = v_1$, $a_j = v_2$, $R_{a_j} = 0$, $a_j = v_1 \wedge R_{a_j} = 0$, $a_j = v_2 \wedge R_{a_j} = 0$

6

We generate MAR data by amputing *hindsight* based on *EURef16* values: individuals with *EURef16 = remain* are more likely to have missing *hindsight* values than individuals with *EURef16 = leave*. Hence, the missing data is MAR since the information about the missing values is observed. Consequently, as expected, the top-1 subgroup is selected using *EURef16 = remain* (cf. Table 6.5). Because the two attributes do not perfectly correlate, a MAR mechanism may ampute *hindsight = right* as well. It has been shown that MAR converges to MCAR for small correlations, and to MNAR for strong correlations [179]. Accordingly, for small missingness percentages, we find that *hindsight* again appears in the subgroups' descriptions.

With MNAR, the information about the missing values is really missing from the data. Therefore, for high missingness percentages, EMM-RCS needs to resort to Definition 6.2 to find the subgroup of non-leavers (cf. Table 6.5, right column, bottom two rows). For medium percentages, the subgroup of missing values is too small to compete with other subgroups: other attributes will be used for describing subgroups.

Missingness in the HBSC and DNSSSU datasets

Next, we experiment with an incomplete dataset, originating from the HBSC [192] and DNSSSU [161] studies. More information about these studies and the deployment of

Table 6.7: Quality and description of the best exceptional subgroup found in the HBSC and DNSSSU datasets with φ_{RCS} for four missing data methods.

Method	φ_{RCS}	**Description**
cca	44.3	age: 12 ∧ urbanity: at least moderate ∧ skipped classes: 0
ignore	46.3	age: 12 ∧ skipped classes: 0 ∧ urbanity: at least moderate
ignoreORselect	46.3	age: 12 ∧ skipped classes: 0 ∧ urbanity: at least moderate
ignoreANDselect	41.9	age: 12 ∧ skipped classes: 0, NaN ∧ sex: girl, NaN

EMM-RCS in this context will be discussed in Section 7.1. In this chapter, we further evaluate the refinement strategy from Definition 6.2.

In particular, we refer to Definition 6.2 as the 'raw' or *ignoreORselect* approach (i.e., individuals with missing values are either ignored or specifically selected) and add an extra approach which we call the *ignoreANDselect* approach: individuals with missing values can be selected together with individuals with observed values. We hypothesize that the *ignoreANDselect* refinement strategy could be beneficial in situations with smaller missingness proportions since observed and unobserved descriptive space can be covered simultaneously. See Table 6.6 for a short description and example selection conditions for the two approaches.

We compare our *ignoreORselect* and *ignoreANDselect* methods with two traditional methods. First, we consider complete case analysis (*cca*), which discards incomplete individuals from the dataset, removing the need to make the refinement strategy handle missing data. The second way of handling incomplete descriptors is by simply *ignoring* individuals with missing values on a particular attribute a_j. These individuals cannot be covered by the description.

Table 6.7 lists the descriptions and qualities of the top-1 subgroup for each approach. Not surprisingly, the top description for *ignore* is similar as the top description for *ignoreORselect*; we expect our data to be Missing Completely At Random (MCAR) and our missingness percentages are small. In contrast, with *ignoreANDselect*, the subgroup set contains many conditions that cover both observed and unobserved individuals, such as *skipped classes: 0, NaN*.

Although all descriptions generated with *ignore* and *ignoreORselect* are theoretically also possible with *ignoreANDselect*, the quality of the top subgroup is lower for the latter. This illustrates a property of beam search parameter w: if we increase the space of candidate subgroup descriptions, at earlier search levels important precursors may be pushed out of the beam, which could lead to reduced quality in the final result set. Increasing the beam width w may resolve this issue; indeed, when we set $w = 60$ rather than $w = 20$, the top subgroup with quality 41.9 drops to position 10.

For *cca*, differences with *ignore* and *ignoreORselect* are subtle; some subgroup descriptions are slightly more general, and the order of some conditions is reversed. The reason is a slight change in the proportions of the observed values.

6.8 Conclusion

We propose Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS): an EMM instance discovering subgroups with exceptional trends in data collected with a Repeated Cross-Sectional (RCS) research design. We develop an expressive quality measure, φ_{RCS}, that builds on the standard error of trend estimates and is easily adapted for finding a variety of exceptionalities. EMM-RCS can handle practical RCS data problems including uneven spacing of measurement over time, fluctuating sample sizes, and incomplete descriptive attributes.

We evaluate the sensitivity of our quality measure through a synthetic data experiment. Our results clearly display the trade-off between the distance, the uncertainty of the trend estimate and the subgroup size in determining the exceptionality of a subgroup. After all, the larger the quality value, the more exceptional the subgroup. A subgroup that contains 25% of the individuals and has a trend line at distance $dist = 1$ with standard deviation $sd = 1$, has the same quality value as a subgroup that contains 25% of the individuals and has distance $dist = 3$ and standard deviation $sd = 3$. Indeed, even though the latter subgroup is further away from the global trend, the uncertainty is larger. Therefore, its exceptionality cannot be distinguished from a subgroup with a trend that is closer but has a smaller uncertainty.

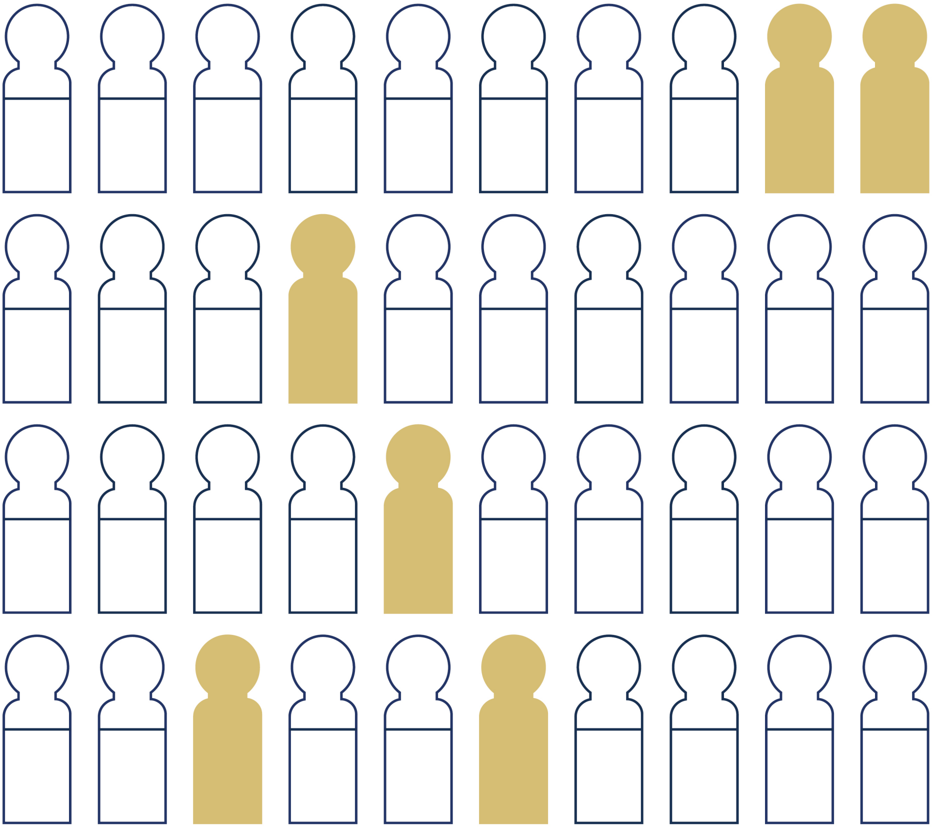
We account for missing data in descriptive attributes by the refinement introduced in Definition 6.2; in Section 6.7.3, we demonstrate its working for MCAR, MAR, and MNAR missingness. We find that with MNAR, the information about the missing values is really missing from the data and Definition 6.2 can be used to discover the true subgroup. For MCAR and MAR, the search algorithm tends to use other descriptive attributes that correlate well with the incomplete descriptor.

Alternatively, the *ignoreANDselect* approach partly solves issues with small missingness percentages, but it is less interpretable. Furthermore, we find that this refinement strategy increases the space of candidate subgroup descriptors in such a way that at earlier search levels, important precursors may be pushed out of the beam. This could effectively lead to a reduced quality in the final result set.

Future extensions to Definition 6.2 or alternative missing data methods for handling incomplete descriptors should be explored. A possible direction for future research could lie in adapting EMM-RCS to handle different sampling designs, including random, stratified and cluster-based sampling. This can be achieved by including weights in the population estimate in Equation (6.3).

Finally, perhaps the starkest illustration of the versatility of EMM-RCS and our quality measure is provided by the results in Figure 6.3, on the Brexit dataset. When looking for groups with an exceptional slope in the trend, we find three subgroups that each show a drastic reduction in identification with the *leave* camp, when comparing measurement occasions directly before and after the landslide victory of Boris Johnson in the 2019 UK General Election. These subgroups share the characteristic of finding that the government handled the negotiations badly, which explains why they are unenthusiastic of Boris Johnson taking a firmer grip on proceedings. But crucially, these subgroups differ in their relation with the overall population: subgroup 2 skewed more, subgroup 1 skewed comparable, and subgroup 9 skewed less than average towards *leave*. Hence, EMM-RCS can detect a change in signal of this kind, independent of how the subgroup behaved relative to the overall population *outside* of this change itself: it can detect both the relatively crude global signals of the dashed line and the relatively subtle local signals of the solid lines.

7



Discovering Subgroups with Exceptional Trend Behavior

Over the last two decades, alcohol use has been in decline among Dutch adolescents. However, the declining trend has been flatlining: prevalence of monthly alcohol use among Dutch 12-to-16-year-olds decreased from 54% in 2003 to 26% in 2013, but merely to 23% in 2019. Dutch governmental policy makers aim to decrease this prevalence further. To do so effectively, it would benefit them to know whether social group memberships correspond to exceptional alcohol use trends. With traditional statistical approaches, it is challenging to analyze such a relation between societal trends and social group memberships: only a few socio-demographic variables can be included, subgroups must be pre-defined, and linearity assumptions are required. We resolve these issues and automatically identify social subgroups of the Dutch adolescent population by deploying Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) on data that interleaves two quadrennial studies: the Health Behaviour in School-Aged Children study (HBSC), and the Dutch National School Survey on Substance Use (DNSSSU). Our findings confirm existing knowledge that age, educational level, and migration background are important descriptors of monthly alcohol use, and provide further insights into the existence of an interplay effect with life satisfaction, urbanization degree, and truancy.

The contents of this chapter have previously appeared in **Schouten, R. M., Stevens, G. W., van Dorsselaer, S. A., Duinhof, E. L., Monshouwer, K., Pechenizkiy, M., and Duivesteijn, W.** Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining: Discovering exceptional trends in adolescent alcohol use in the Netherlands. *Accepted for presentation at BNAIC/BeNeLearn* (2024) [177].

7.1 Introduction

Analyzing societal trends is an important line of research in social sciences, as it assesses how the behaviors, attitudes, and feelings of populations change over periods of time, and for which groups such changes are particularly pronounced. Such insights are not only scientifically important, but also have the potential to pinpoint directions for policies and interventions. For instance, the European School Survey Project on Alcohol and Other Drugs (ESPAD) collects data on substance use and other forms of risk behavior among 15- to 16-year-old students in 49 European countries [137], and Monitoring the Future focuses on drug and alcohol use among students in America [94]. Using data of the Health Behaviour in School-Aged Children study (HBSC) [192] and the Dutch National School Survey on Substance Use (DNSSSU) [161], we analyze trends in adolescent alcohol use between 2003 and 2019 in the Netherlands. We aim to demonstrate the value of deploying a local pattern mining approach as a sociological method by identifying subgroups of adolescents displaying deviating patterns of alcohol use.

Lifetime and monthly alcohol use among Dutch adolescents has changed dramatically over the last decades: a substantial increase between 1992 and 2003 [40] was followed by a sharp decrease between 2003 and 2015. Since 2015, both lifetime and monthly alcohol use among Dutch adolescents has remained stable [161, 192]. The monthly prevalence of alcohol use in 2019 still ranges from 5% among 12-year-olds to 53% among 16-year-olds [161]. Also, there are sizable subgroup differences in alcohol use. Higher prevalence of lifetime alcohol use can be found for older adolescents versus younger adolescents, adolescents with lower versus higher educational levels, adolescents without a migration background versus those with a migration background, and adolescents from families with a relatively high versus low socioeconomic status [192].

Science is well aware that early and frequent alcohol use leads to a broad range of negative consequences [133, 186]. Policy makers, government institutions, and other decision makers are interested in further reducing alcohol use among adolescents. To that end, it would help to have a better understanding of factors that influence when, whether, and why the downward trend flatlines. This calls for a structured search through a space of demographic subgroups, and evaluation of their exceptionality in terms of behavior of a response variable in repeated cross-sectional data (which is the form in which data from both the HBSC and DNSSSU studies arrives). Hence, in this chapter, we deploy Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) [171] (**Chapter 6**) as a sociological method.

EMM-RCS falls under the umbrella of Exceptional Model Mining (EMM) [54], which generally searches for coherent subgroups in a dataset that behave somehow exceptionally. In EMM-RCS, this behavior becomes a societal trend that deviates from the average, population trend. Any type of deviation in the alcohol prevalence or trend thereof could be helpful for developing tailored interventions and policies. Schools would want to know how life satisfaction relates to trends in adolescent alcohol use. For public health departments, the relation with the degree of urbanization can be very informative. Furthermore, it would be relevant to discover for which groups trends in monthly alcohol use follow a different course than the average, population trend (i.e., a stronger or weaker decrease over time), or trends could run fairly stable. This might indicate subgroups of adolescents that are exceptionally sensitive or insensitive to certain policy measures. In this chapter, we seek demographic subgroups displaying the following three types of exceptional trend behavior: (1) deviations of the prevalence of monthly alcohol use; (2) deviations in the course (i.e., change-over-time) of the trend in monthly alcohol use; (3) horizontal trends (i.e., no change-over-time) in monthly alcohol use.

7.2 Related sociological work

Analyzing the extent to which trends in adolescent alcohol use in the last decades vary across subgroups is challenging for several reasons. First, with traditional statistical approaches, it is difficult to include many socio-demographic variables in the analysis because of the risk of an increased type-I error rate due to multiple hypothesis testing. In order to prevent the finding of false significant results, it is therefore common to select a few socio-demographic variables based on theory or existing literature. Although it is sensible to restrict the number of statistical tests, in this way the possibility to discover new types of subgroup-specific trends is limited. For instance, trends in adolescent alcohol use are mostly analyzed for subgroups based on gender, age, and educational track [39, 40, 68, 160], but other variables such as ethnic background and family situation are rarely included in the analysis or merely used as covariates.

The constraint on the number of tests also complicates the evaluation of combinations of socio-demographic variables. According to intersectionality theory [36], adolescents belong to multiple social groups and their social experiences are shaped by all these social group memberships together. Subsequently, the effect of belonging to a particular social group on adolescents' developmental outcomes should not be considered separately from other social group memberships [35, 69]. When investigating the extent to which adolescent alcohol use vary across subgroups, a common approach to studying the interplay between social group memberships is to create dummy variables that indicate group membership of combinations of socio-demographic variables. This can be done by performing multiple group logistic regressions by creating 6 dummy variables for the combined membership of a social group based on gender (2 groups) and age (3 groups) [39], or with a structural equation model [68]. Apart from the fact that the manual discretization of a continuous variable into groups (as is done here for the variable age) may already limit the potential for finding interesting interactions with other variables, a statistical approach where every subgroup is a separate dummy variable greatly reduces the number of group memberships that can be considered. After all, the number of combinations of group memberships scales exponentially with the number of socio-demographic variables.

A further complexity in analyzing the extent to which societal trends vary across subgroups is the repeated cross-sectional research design that is used to collect trend data [25]. For instance, the change-over-time of the monthly prevalence of alcohol use is assessed by repeatedly sampling new individuals from a population at successive measurement moments. This restricts statistical analysis to models that analyze changes in the trend with respect to a reference year, by dummy-coding survey year, instead of considering the variation over time as a continuous event as can be done in longitudinal or time series data. When a regression model is used, survey year can be added as an independent, continuous variable but its relationship with other independent variables or with the (log-odds of the) dependent variable requires the assumption of linearity.

7.3 Data collection

Adolescent alcohol use in the Netherlands is monitored by two studies: the Dutch National School Survey on Substance Use (DNSSSU) [161] and the Health Behaviour in School-aged Children study (HBSC) [192]. Both studies are conducted every four years, with an offset of two years between them: combining both studies results in 2-yearly data. We use DNSSSU data from 2003, 2007, 2011, 2015, and 2019, and HBSC data from 2005, 2009, 2013, and 2017. For both studies, we use data of students in the first four years of secondary school aged 12 to 16 (excluding younger or older students, as their numbers are very small). Data are collected in October and November with a two-stage random sampling design where schools are stratified by region. Within each school, 2 to 5 classes are randomly selected to participate, and we collect data from all students in these classes. The adolescent response rate in the classroom was above 92% in all years. Non-response on the student level mainly occurs because of illness. The two studies use self-complete paper-and-pencil questionnaires from 2003 to 2013 and computer-assisted questionnaires from 2015 onward.

Both HBSC and DNSSSU assess alcohol use by asking adolescents how often they drank alcohol in their entire life, in the last 12 months, and in the last 4 weeks. In this chapter we use data on the 4-week prevalence. From 2003 to 2011, answer options were 0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11–19, 20–39, and 40 or more. The question has been asked in a subtly different form from 2013 onwards: it asks for the number of days that adolescents drank alcohol, with answer options never, 1–2 days, 3–5 days, 6–9 days, 10–19 days, and 30 days or more. For this study, we flatten answer options 0 (2003–2011) and never (2013–2019) into 0, and the other answer options into 1, resulting in the monthly prevalence of alcohol use.

We include in the analysis all socio-demographic variables that were available for all waves, resulting in 10 variables: 2 are continuous (age, life satisfaction), 2 are dummy-coded (sex, whether the adolescent lives with both parents), 3 are nominal (ethnic group, whether father has a job, whether mother has a job), and 3 are ordinal (school level, level of urbanity, number of skipped classes/truancy). We let missing values be, since EMM-RCS can natively handle missingness in socio-demographic variables [171].

7.4 EMM-RCS deployment

Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) was introduced in Section 6.4 as a generic method to discover subgroups displaying exceptional trending behavior across waves in repeated cross-sectional data. Here, we deploy EMM-RCS as a sociological method, in order to understand the relation between socio-demographic variables and trends in alcohol use.

Candidate subgroups are formed using combinations of selection conditions on socio-demographic variables; these select adolescents from the full population. Such a combination is called a *description*, and could be as follows: *age = 12 ∧ life satisfaction 8-10 ∧ 0 skipped classes*. The number of possible combinations of social group memberships is explosive, and evaluating the exceptionality for every candidate subgroup is infeasible (on top of which, it would severely increase the type-I error rate). A search strategy is necessary to efficiently traverse this space; our choice, out of the many strategies that exist for this purpose, is detailed in Section 7.4.2.

7.4.1 Instantiations of our quality measure

For each generated subgroup, we must determine how exceptional its trend is. This is captured through the definition of a quality measure φ, where canonically, higher values represent higher exceptionality of behavior.

We denote data collected with a repeated cross-sectional research (RCS) design as follows: $\Psi = (\Omega_{x_1}, \ldots, \Omega_{x_t}, \ldots, \Omega_{x_T})$ is an ordered bag of T datasets where each Ω_{x_t} is collected at *wave* x_t for $x_t \in \mathcal{T}$. Since we analyze trends in alcohol use in the Netherlands between 2003 and 2019, $\mathcal{T} = \{2003, 2005, ..., 2019\}$ and $t \in \{1, 2, ..., T\}$ with $T = 9$. Every Ω_{x_t} is a bag of records (i.e., individuals, adolescents) $r_{x_t} \in \Omega_{x_t}$ of the form $r_{x_t} = (a_1, \ldots, a_k, \ell_{x_t})$, where $a_1, ..., a_k$ are the sampled values from k socio-demographic variables and $\ell^i_{x_t} \in \{0,1\}$ is a binary value indicating whether adolescent $r^i_{x_t}$ has drunk alcohol in the past month. Following statistical theory [22], if variable ℓ_{x_t} has a binomial distribution with parameters n_{x_t} and μ_{x_t} and when n_{x_t} is large, μ_{x_t} can be approximated by the proportion of the sampled values ℓ_{x_t} with associated standard error $se(\mu_{x_t})$:

$$\mu_{x_t} = \frac{1}{n_t} \sum_{i=1}^{n_t} \ell^i_{x_t}. \tag{7.1}$$

$$se(\mu_{x_t}) = \sqrt{\frac{\mu_{x_t}(1-\mu_{x_t})}{n_{x_t}-1}}. \tag{7.2}$$

We say μ_{x_t} is the prevalence of alcohol use at wave x_t. The trend in monthly alcohol use is then the collection of the prevalence estimates of all waves together: $\{\mu_{2003}, \mu_{2005}, ..., \mu_{2019}\}$. Note that since Ψ is an RCS dataset, the sample sizes may differ per wave: we may have $n_{x_t} \neq n_{x_{t'}}$, $t \neq t'$. The total dataset size is $n^\Psi = \sum_{t=1}^T n_{x_t}$.

To gauge whether a subgroup's trend in monthly alcohol use deviates from the population trend, EMM-RCS uses a generic quality measure [171]:

$$\varphi_{\text{RCS}}(D) = f\left(\{z_{x_t} \mid x_t \in \mathcal{T}\}\right) \tag{7.3}$$

$$z_{x_t} = \frac{|\theta^{SG}_{x_t} - \theta^0_{x_t}|}{se(\theta^{SG}_{x_t})}. \tag{7.4}$$

In the following, we detail how to instantiate Equations (7.3) and (7.4) to measure the three types of exceptionality that are relevant for better understanding adolescent alcohol use, as outlined at the end of Section 7.1.

Exceptional deviations of the prevalence

In order to discover subgroups of adolescents with trends in monthly alcohol use that deviate from the average, population trend at any, unknown wave, we define $\theta^{SG}_{x_t} = \mu^{SG}_{x_t}$

and $\theta^{0}_{x_t} = \mu^{\Psi}_{x_t}$ for all $x_t \in \mathscr{T}$. In other words, as a statistic in Equation (7.4) we use the monthly prevalence of alcohol use as defined in Equation (7.1). Superscripts SG and Ψ refer to the subgroup and the entire dataset respectively. Consequently, for every wave, we compare the prevalence of monthly alcohol use in the subgroup with the prevalence in the entire dataset.

We furthermore set $f(\cdot) = \max$, which means that for a given subgroup, we select the maximum of the T z-scores. In other words, the largest difference between a subgroup's trend estimates and the average, population trend estimates serves as the exceptionality value of the subgroup. In practice, this means that we could both select a subgroup with a trend that is very similar to the population trend but suddenly deviates at one particular wave, and a subgroup with a trend that deviates over the entire course.

Exceptional slope deviations

Subgroups with trends with an exceptional increase or decrease can be discovered by focusing on the slopes of the prevalence estimates. A slope is simply the difference between two subsequent prevalence estimates. In order to account for small fluctuations between the prevalence values estimated in HBSC and DNSSSU, we first take a weighted moving average of two subsequent prevalence estimates, and then calculate the slope between two averages. Denoting a weighted moving average of the prevalence estimates at occasions x_t and x_{t-1} with τ_{x_t}, we define the slope as $\theta_{x_t} = \tau_{x_t} - \tau_{x_{t-1}}$. Note that for T waves, there will be $T-1$ averages and $T-2$ slopes. The standard error of the weighted average of two proportions follows from statistical theory,

$$se(\tau_{x_t}) = \sqrt{b^2_{x_t} se(\mu_{x_t})^2 + b^2_{x_{t-1}} se(\mu_{x_{t-1}})^2}, \tag{7.5}$$

where the weights b are based on the sample sizes n_{x_t} and $n_{x_{t-1}}$. The standard error of the slope is similar but with weights $b = 1$.

We compare the slopes of the subgroup's trend with the slopes of the population trend. Therefore, $\theta^{0}_{x_t} = \theta^{\Psi}_{x_t}$. Again, we choose $f(\cdot) = \max$, which means that we consider subgroups to be exceptional when there is some slope at some wave that greatly differs from the slope in the population trend. We could thus discover subgroups with a sudden increase or decrease in the trend and subgroups with completely deviating courses.

Exceptionally horizontal trends

Discovering subgroups of adolescents with horizontal trends in monthly alcohol use does not require a comparison between a subgroup's trend and the average, population trend. Therefore, $\theta^{0}_{x_t} = 0$. Furthermore, we define $\theta^{SG}_{x_t} = \tau_{x_t} - \tau_{x_{t-1}}$ to be the slope of the weighted moving average in the subgroup. In order to directly evaluate whether the slope estimate is close to 0, we define $se(\theta^{SG}_{x_t}) = 1$. Then, we first select all the slopes that are close to zero with a certain threshold ϵ and sum the absolute difference between these slopes and that threshold ϵ. Formally, $f(\cdot) = f_{\text{countsum}}(\epsilon) = \sum\{\text{abs}(z_{x_t} - \epsilon) \mid x_t \in \mathscr{T}'', z_{x_t} < \epsilon\}$ with $|\mathscr{T}''| = T-2$ because T waves give $T-1$ weighted moving averages and $T-2$ slopes. In this way, we favor subgroups that have many slopes that are close to zero (because the count will be high) and distinguish between subgroups by favoring the most horizontal ones (because the absolute difference will be high). The value for ϵ can be chosen based on theory or by means of one of the validation methods; we report parameter sensitivity experiments for ϵ in Section 7.5.3.

7.4.2 Experimental approach

Table 7.1 provides an overview of the sample size, monthly prevalence of alcohol use and associated standard error per year for the entire dataset. Overall, the trend in alcohol shows a linear decrease from 2003 to 2015 and a stable pattern from 2015 onward. The population trend can also be found as the black trend line in all subfigures in Figure 7.1.

To discover exceptional subgroups of adolescents, we employ *beam search*, with parameters $w = 20$, $d = 3$, and $q = 20$ (see Algorithm 1 in Section 2.3.2). Since the number of socio-demographic variables in this study is limited ($k = 10$), we expect that higher settings will lead to spurious subgroups through beam pollution and reduced interpretability of results. We dynamically discretize continuous variables using the `lbca` strategy with octiles from [135].

In particular in this chapter, we evaluate three techniques that validate the significance of discovered subgroups:

1. we employ Dominance-Based Pruning (DBP) [201], removing conditions that decrease the quality of a subgroup,
2. we let resulting subgroups undergo validation with the null Distribution of False Discoveries (DFD) [55] with parameters $m = 100$ and a one-sided significance level $\alpha_{DFD} = 0.025$,
3. we demand that added conditions improve the subgroup quality over the parent node in the beam by at least 2.5%. This Minimum Improvement (MI) threshold was determined together with domain experts: an increase in the proportion of adolescents that drank alcohol in the last month should be ≥ 1 percentage point.

Furthermore, we aim to reduce redundancy in the result list using a Weighted Coverage Scheme (WCS) ($\gamma = 0.9$) and Description-Based Selection (DBS) with a fixed size of $2w$ (see Section 2.3.3 for more background information).

7.5 Discovering exceptional trends in alcohol use

In the following, we discuss the demographic subgroups that we found with each of the three types of exceptionality. In each subsection, we report a table (Tables 7.2–7.4) of all subgroups found *before* pruning. The subgroups are manually split into Trend Groups (TG) that share certain characteristics. From each TG, we take the most exceptional subgroup and display its trend in the corresponding subfigure of Figure 7.1. Trends of the other subgroups, *after* pruning, are available on our interactive dashboard.[1]

[1] We provide a link to the dashboard and additional material such as descriptive information, missingness percentages per category per variable per year, information on data availability, and all experimental code at `https://github.com/RianneSchouten/AlcoholTrends_HBSCDNSSSU_EMM/`.

7.5.1 Discovering exceptional deviations of the prevalence

In this section, we present results for social group memberships that lead to deviations of the prevalence of monthly alcohol use between 2003 and 2019 in the Netherlands. The original top-$q = 20$ discovered by beam search before the application of any pruning strategies is listed in Table 7.2. DBP and validation with the DFD have not resulted in any changes in the top-20 subgroups of adolescents. The false discoveries were distributed around a mean quality of 4.22 (SD: 0.48) which results in a threshold value at α_{DFD} of 5.17. Because the subgroup with the lowest quality value has a value of 25.6, none of the top-20 subgroups are rejected. Validation with MI removes 5 conditions on socio-demographic variables. For three subgroups, the new descriptions (after removing a condition) were similar to an existing description in the top-20, which reduced the list of subgroups to a top-17.

Comparing the prevalence of adolescent alcohol use in the past month in Figure 7.1a with the overall population, prevalence is down in two trend groups (1 and 2) and up in the other three (3, 4, and 5). The subgroups with trends below the population trend describe adolescents who are relatively young, who do not skip classes, who are fairly satisfied with their life, and who live in moderately to highly urbanized areas. In the trend groups where more adolescents drink alcohol, age is an important factor as well. Here, relatively older adolescents are selected. Being in an older age group interacts with having a Dutch or western ethnicity (thus excluding non-western ethnicities).

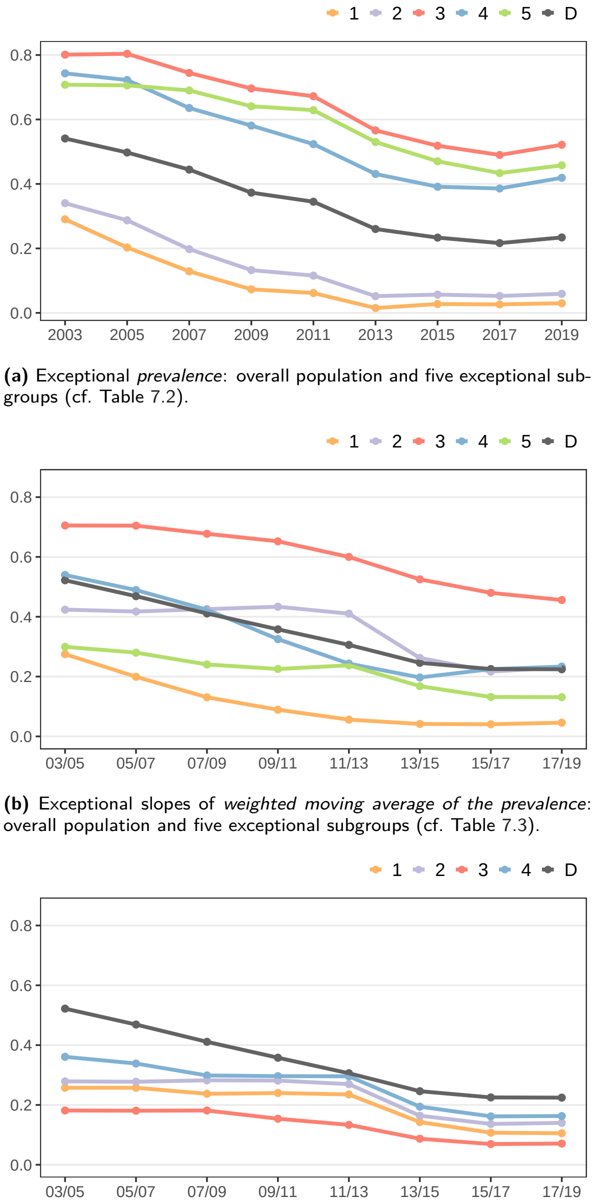
For all trend groups, the entire trend line deviates from the population trend. Even though the quality measure focuses on a maximum deviation at *any* point in time (Section 7.4.1), the prevalence in monthly alcohol use differs from the population prevalence at *all* waves (Figure 7.1a). Apparently, a subgroup trend that deviates from the population trend at one particular wave likely deviates at other occasions too. The change-over-time of alcohol use resembles that of the population trend in all subgroups, maybe except for trend group 5 (subgroup 13); monthly alcohol use decreases between 2003 and 2015 and flattens afterwards.

7.5.2 Discovering exceptional slope deviations

This section presents results for social group memberships that lead to deviations in the course (i.e., change-over-time) of the trend in monthly alcohol use. The original top-$q =$ 20 discovered by beam search before the application of any pruning strategies is listed in Table 7.3. DBP and validation with the DFD (mean quality 2.55 with SD of 0.21, value at α_{DFD} is 3.07, smallest quality value is 3.8) does not result in any changes in the top-20

Table 7.1: Sample size, prevalence estimate and associated standard error per year in the entire dataset.

Survey	DNSSSU	HBSC	DNSSSU	HBSC	DNSSSU	HBSC	DNSSSU	HBSC	DNSSSU
Year	2003	2005	2007	2009	2011	2013	2015	2017	2019
n	6791	5272	6234	5490	6374	5421	6232	6060	5022
PREV	0.54	0.50	0.44	0.37	0.34	0.26	0.23	0.22	0.23
se(PREV)	0.0061	0.0069	0.0063	0.0065	0.0060	0.0060	0.0054	0.0053	0.0060



(a) Exceptional *prevalence*: overall population and five exceptional subgroups (cf. Table 7.2).

(b) Exceptional slopes of *weighted moving average of the prevalence*: overall population and five exceptional subgroups (cf. Table 7.3).

(c) Exceptionally horizontal *weighted moving average of the prevalence*: overall population and four exceptional subgroups (cf. Table 7.4).

Figure 7.1: Exceptional trends in monthly alcohol use among Dutch adolescents for three types of trend deviations in monthly alcohol use: (a) exceptional deviations of the prevalence; (b) exceptional deviations in the course of the trend; (c) exceptionally horizontal trends. Black displays the population trend.

Table 7.2: Top-20 subgroups of adolescents with exceptional deviations of the prevalence of monthly alcohol use. Validation with a minimum improvement threshold results in the removal of 5 conditions (in red; the quality improvement is 1.4, 1.7, 0.4, -1.0 and 0.3 percent respectively). Three conditions narrowly exceed the threshold with 2.6, 2.8, and 2.6 percent, respectively (in orange).

TG	SG	Cov	Description
1	1	0.11	age: 12 ∧ skipped classes: 0 ∧ urbanity: at least moderate
	2	0.15	age: 12 ∧ life satisf: 7-10 ∧ skipped classes: 0
	3	0.14	age: 12 ∧ life satisf: 7-10 ∧ urbanity: at least little
	10	0.09	age: 12 ∧ skipped classes: 0 ∧ sex: girl
2	4	0.35	age: 12-13 ∧ skipped classes: 0 ∧ life satisf: 7-10
	5	0.37	age: 12-13 ∧ skipped classes: 0 ∧ life satisf: 6-10
	7	0.40	age: 12-13 ∧ skipped classes: 0
	9	0.25	age: 12-13 ∧ life satisf: 6-10 ∧ urbanity: at least moderate
	12	0.37	age: 12-13 ∧ life satisf: 7-10
	14	0.40	age: 12-13 ∧ life satisf: 6-10
	16	0.41	age: 12-13 ∧ skipped classes: 0-1
	18	0.43	age: 12-13
	20	0.34	age: 12-13 ∧ school level: at least vmbo-t
3	6	0.26	age: 15-16 ∧ ethnicity: dutch, western
	8	0.24	age: 15-16 ∧ ethnicity: dutch
4	11	0.48	age: 14-16 ∧ ethnicity: dutch, western
	15	0.44	age: 14-16 ∧ ethnicity: dutch
5	13	0.32	age: 15-16
	17	0.29	age: 15-16 ∧ life satisf: 0-9
	19	0.29	age: 15-16 ∧ father job: yes, don't know

subgroups of adolescents with exceptional trend deviations. Validation with MI indicates the removal of 8 conditions from 7 subgroups; 6 of which have a description that is similar to another one. 14 subgroups will be remaining.

Figure 7.1b presents the weighted moving average of the trends in monthly alcohol use for five main trend groups. Note that not all subgroups in the same trend group have similar trend values (as was the case in Section 7.5.1). For instance, while subgroups 11 and 15 are assigned to trend group 1, they are much larger than subgroups 1, 4, 5, and 8, and hence their trend in alcohol use will be closer to the population trend (higher prevalence values). However, the trend courses (i.e., change-over-time) of subgroups 11 and 15 are similar to those of the other subgroups in trend group 1.

We find various trend shapes. Trend groups 1 and 3 (red and orange lines in Figure 7.1b) decrease at a slower pace than the overall population, while trend group 4 (blue line) decreases much faster, especially between 2009 and 2015. Between 2003 and 2013, the slopes of trend groups 2 and 5 (purple and green lines) are fairly close to 0, while the population trend decreases in those years.

The variation in exceptional trend courses is also reflected in the corresponding social group memberships. In trend group 1, adolescents who are young and satisfied with life are less likely to drink alcohol than the average adolescent. Trend group 3 has a low de-

Table 7.3: Top-20 subgroups of adolescents with exceptional deviations in the course of the trend in monthly alcohol use (i.e., exceptional slope deviations). Validation with a minimum improvement threshold results in the removal of 8 conditions (in red, the quality improvement is -1.4, -2.9, -1.4, -1.5, -4.3, 2.4, 2.1, and -2.3 percent, respectively).

TG	SG	Cov	Description
1	1	0.19	age: 12
	2	0.17	age: 12 ∧ ethnicity: dutch, non-western
	3	0.15	age: 12 ∧ complete family: yes
	4	0.17	age: 12 ∧ ethnicity: dutch, non-western ∧ life satisf: 0-10
	7	0.16	age: 12 ∧ ethnicity: dutch, western
	8	0.15	age: 12-13 ∧ life satisf: 9-10
	9	0.29	age: 12-13 ∧ life satisf: 8-10
	13	0.37	age: 12-13 ∧ life satisf: 7-10
	17	0.41	age: 12-13 ∧ ethnicity: dutch, non-western
	15	0.54	age: 12-14 ∧ ethnicity: dutch
	20	0.53	age: 12-14 ∧ mother job: yes
2	5	0.1	urbanity: very high ∧ age: 14-16
	11	0.14	urbanity: very high ∧ life satisf: 0-9
	19	0.14	urbanity: very high ∧ age: 13-16
3	6	0.26	age: 15-16 ∧ school level: at least vmbo-p/t
	14	0.32	age: 15-16
	18	0.11	skipped classes: ≥ 1
4	10	0.23	school level: at least havo ∧ urbanity: at most moderate
5	12	0.14	age: 15-16 ∧ urbanity: at least high
6	16	0.13	school level: vmbo-p/t - havo ∧ ethnicity: western, non-western

crease in alcohol use as well, but this concerns older adolescents (subgroups 3 and 9) or adolescents who skip classes in school (subgroup 12). The conditions on age and the condition on skipped classes do not appear in a description together: they describe two distinct social group memberships. This could indicate that these factors singularly induce an effect on alcohol use and that there is no interplay.

In trend group 2, the urbanization degree of the area where adolescents live has an important relation with the trend in alcohol use. In 2003, about 40% of the adolescents who live in highly urbanized areas had drunk alcohol in the past month. That percentage stayed stable until 2013, when it suddenly dropped to about 20%. Similar effects occur in trend group 5 (subgroup 11), encompassing adolescents in lower educational tracks with a non-Dutch ethnicity. An inverted trend course is followed by trend group 4 (subgroup 6). Between 2003 and 2013 the number of adolescents who drank alcohol has decreased stronger than in the overall population, and that decrease suddenly stopped around 2013. Here, we find an interplay between a school level that is at least HAVO and a living area that is at least moderately urbanized.

7.5.3 Discovering exceptionally horizontal trends

This section presents results for social group memberships that lead to exceptionally horizontal trends in monthly alcohol use. The original top-$q = 20$ discovered by beam search

Table 7.4: Top-20 subgroups of adolescents with exceptionally horizontal trends in monthly alcohol use. Validation with the DFD results in the removal of 10 subgroups (in red).

TG	SG	Cov	Description
1	1	0.09	ethnicity: non-western ∧ life satisf: 0-8 ∧ skipped classes: 0-2
	3	0.09	ethnicity: non-western ∧ life satisf: 0-8 ∧ skipped classes: 0-4
	4	0.08	ethnicity: non-western ∧ life satisf: 0-8 ∧ school level: ≤ havo/vwo
	6	0.08	ethnicity: non-western ∧ age: 13-16 ∧ school level: ≥ vmbo-t
2	2	0.08	ethnicity: non-western ∧ age: 14-16 ∧ urbanity: ≥ moderate
	15	0.09	ethnicity: non-western ∧ age: 14-16 ∧ urbanity: ≥ little
	20	0.08	ethnicity: non-western ∧ age: 14-16 ∧ skipped classes: 0-2
3	5	0.10	ethnicity: non-western ∧ complete family: yes ∧ skipped classes: 0-4
	11	0.10	ethnicity: non-western ∧ complete family: yes ∧ father job: yes,no
4	9	0.11	school level: ≥ havo/vwo ∧ ethnicity: (non)-western ∧ life satisf: 0-8
	7	0.28	age: 12-13 ∧ life satisf: 8-10 ∧ skipped classes: 0-1
	8	0.07	skipped classes: ≥ 1 ∧ age: 14-16 ∧ school level: ≤ havo/vwo
	10	0.14	age: 12 ∧ ethnicity: dutch ∧ skipped classes: 0-6
	12	0.30	age: 12-13 ∧ life satisf: 7-10 ∧ school level: at least vmbo-t
	13	0.08	urbanity: very high ∧ school level: ≤ havo ∧ age: 13-16
	14	0.17	age: 12-13 ∧ sex: boy ∧ school level: ≤ havo/vwo
	16	0.08	age: 12-13 ∧ life satisf: 9-10 ∧ school level: ≤ havo
	17	0.10	age: 12 ∧ life satisf: 6-8
	18	0.36	age: 12-13 ∧ life satisf: 7-10 ∧ skipped classes: 0-1
	19	0.10	sex: girl ∧ ethnicity: (non)-western ∧ skipped classes: 0-4

before the application of any pruning strategies is listed in Table 7.4. When ϵ is set to the prevalence points 0.005, 0.01, and 0.02, validation with the DFD results in the rejection of 16, 10, and 3 subgroups, respectively (out of $q = 20$ subgroups). In other words, when we set the threshold too strictly ($\epsilon = 0.005$) the found subgroups are spurious results. However, when we set the threshold too loosely ($\epsilon = 0.02$), we find subgroups with trends that are not horizontal at all (for $T = 9$ waves, a decrease of 0.02 prevalence point per occasion would allow the prevalence to drop with 0.16 over the entire measurement period; it is indeed questionable whether such a trend can be considered flat). Therefore, we apply $\epsilon = 0.01$ and end up with 10 subgroups.

Being a member of a non-western ethnic group is an important factor in all trend groups. The trends in alcohol use are fairly horizontal from 2003 to 2013, then drop, and again stay horizontal after 2015. A similar pattern was found in Section 7.5.2, but there we selected adolescents who live in highly urbanized areas or who attend a lower educational track and have a non-Dutch ethnicity (trend groups 2 and 5 in Table 7.3). When we specifically search for horizontal trends, having a non-western ethnicity turns out be the most dominant factor.

The other conditions on socio-demographic variables in Table 7.4 may confuse because they seem to select the general population of adolescents. For instance, it is likely that most adolescents have a life satisfaction between 0 and 8 and may skip between 0 and 2 classes (subgroup 1). However, all these conditions have passed the minimum improvement threshold. To understand the relevance of conditions 2 and 3 in Table 7.4, it is useful to consider them as exclusion criteria rather than a selection. In the population, adolescents with a non-western ethnicity make up 15%. If, from that group, adolescents are excluded who have a life satisfaction of 9 or 10 and have skipped at least 3 classes, subgroup 1 contains only 9% of the adolescents. Hence, this exclusion is a reduction of 40%. It indicates that adolescents with a non-western ethnicity who additionally have a very high life satisfaction and skip classes do not have such a stable and horizontal trend in alcohol use as adolescents with a non-western ethnicity who have a low to average life satisfaction and skip maximally 2 classes.

A similar reasoning can be applied to the other subgroups; given that an adolescent has a non-western ethnicity, the trend in alcohol use is not as horizontal for those who are relatively young (age 12, 13; subgroups 2, 6, 9, 10) or who do not live with both parents (subgroups 5, 8). A combination with conditions on school level, degree of urbanization, and number of skipped classes may increase this effect further. Overall, having a non-western immigration background is an important factor for having a stable trend in alcohol use, but within this group there are adolescents whose alcohol use is less stable.

7.6 Discussion

Discovered subgroups of adolescents with exceptional trends in monthly alcohol use confirm existing knowledge that for younger adolescents, interactions with memberships of social groups that do not skip classes, have a high life satisfaction and live in moderately to highly urbanized areas lead to a lower prevalence of monthly alcohol use (Section 7.5.1). For older adolescents, we find an interaction with Dutch or western ethnicity that leads to a higher prevalence of monthly alcohol use. Notwithstanding the confirmation that age has a strong relation with alcohol use [161, 192], EMM-RCS discovers interactions with other socio-demographic variables that provide relevant information not only for policy makers but also as a starting point for further research. For example, we discover a relation between ethnic background and having a horizontal trend in monthly alcohol use, both when analyzing general deviations in the course of the trend (Section 7.5.2) and when specifically searching for horizontal trends (Section 7.5.3).

A more in-depth understanding of subgroups of adolescents displaying stable alcohol use trends is important: domain knowledge determines whether such trends are worrisome and warrant attention. Examples of opposite hypotheses can be drawn from subgroups reported in Section 7.5.2: for the adolescents who are young and satisfied with life, the prevalence is already so low that it cannot decrease further and hence the subgroup is not that interesting; for the older adolescents or adolescents who skip classes, the more stable trend may indicate an insensitivity to existing policies and interventions.

Our results spawn two hypotheses for further sociological research. On the one hand, for adolescents with a higher educational level and for adolescents living in at least moderately urbanized areas, the trend in alcohol use decreases at a faster pace (cf. Section 7.5.2). This may indicate that these groups of adolescents are particularly sensitive to policies or interventions. On the other hand, the following social group memberships resulted in a horizontal trend between 2003-2013, followed by a sudden drop in 2013: having a non-western ethnicity and a lower educational track (subgroup 11 in Table 7.3, subgroups 6, 7 in Table 7.4), having a non-western ethnicity and living in an area with a low urbanization degree (subgroups 2, 9 in Table 7.4) and living in an area with a high urbanization degree (subgroups 2, 7, 13 in Table 7.3).

Not for all forms of exceptional trend deviations we find evidence that there is interplay between social group memberships. We find subgroups of older adolescents and a subgroup of adolescents who skip classes whose trend decreases at a slower pace than the population trend (cf. Section 7.5.2). Here, the hypothesis that being a member of multiple social groups has a cumulative [35, 69] or even an aggravated effect on adolescent alcohol use (a.k.a. the multiple jeopardy hypothesis) [49, 100] cannot be accepted. This is in line with other studies that also do not find unequivocal support for the joint effects of subgroup memberships on adolescent mental health [99].

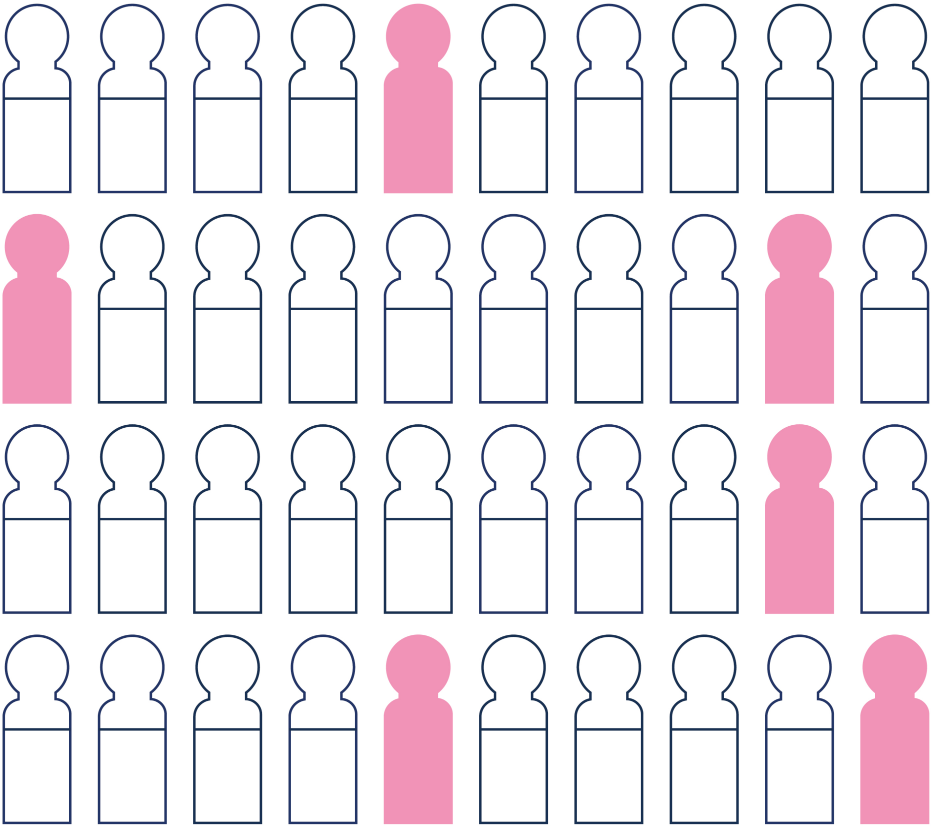
7.7 Conclusion

Analyzing societal trends is an important line of research in social sciences, as it assesses how the behaviors, attitudes, and feelings of populations change over periods of time and for which groups this is particularly true. However, analyzing the extent to which societal trends vary across subgroups in the population is challenging. The number of socio-demographic variables that can be included in the analysis is often restricted. Traditional statistical approaches rely on the assumption of linearity, require manual discretization or dummy-coding, or may not be suitable for data collected with a repeated cross-sectional research design. Hence, it is difficult to study the interplay between social group memberships and to assess whether combinations of conditions on socio-demographic variables have a cumulative or aggravated effect on the course of the trend.

We demonstrate the value of deploying Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) [171] as a sociological method. EMM-RCS uses a heuristically-guided search algorithm that discovers subgroups with trends that deviate from the average, population trend. Because subgroups are formed by selecting people who are a member of particular combinations of socio-demographic groups, EMM-RCS provides interpretable, highly relevant information for policy makers about the needs of specific subgroups in society.

We analyze trends in adolescent alcohol use between 2003 and 2019 in the Netherlands. We investigate whether we can determine (combinations of) social group memberships displaying 1) deviations of the prevalence of monthly alcohol use, 2) deviations in the course of the trend in monthly alcohol use, and 3) exceptionally horizontal trajectories in the trends in monthly alcohol use. Our findings confirm existing knowledge that age, educational level, and migration background are important descriptors of monthly alcohol use, and that an interplay effect exists with life satisfaction, urbanization degree, and truancy. Our findings also spawn two hypotheses for further sociological research, and provide disconfirming evidence to a sociological hypothesis that aligns with existing studies. EMM-RCS thus serves as a hypothesis-generating source that due to its exploratory nature works as a starting point in further understanding the interplay between socio-demographic variables and societal trends.

8



Exceptional Learning Behavior in Descriptive and Target Space

Numerical processing competences such as the ability to enumerate small sets of dots and to compare the relative magnitudes between sets are diagnostic markers of young children's emerging math abilities. In the FUnctional Numerical Assessment (FUNA) study, these abilities are assessed using several computer-assisted tasks, among which is a Dot Enumeration (DE) task where children determine the number of dots in a visual array. It seems that there is a natural threshold around 3 or 4 dots: below this threshold, it is possible to determine the correct number at a glance, known as subitizing; above the threshold, children must count the dots in some way. In this chapter, we develop a piecewise linear regression model class for Exceptional Model Mining with various quality measures discovering subgroups of children whose subitizing curves exhibit atypical patterns. The dataset does not follow the conventional data mining representation where each individual is described with a tuple of attribute values. Rather, for each task, students perform multiple items, one after the other, taken from a larger set of items, and not necessarily in the same order. Hence, we discuss a manner (tailored to the dataset at hand) to transform this item-performance data into the flat-table form that the typical data mining task expects. Domain experts confirm that our experiments evidently demonstrate how children's subitizing performance and counting skills are related to math abilities. Our findings provide opportunity for further development of assessment tools and intervention programs.

The contents of this chapter have previously appeared in **Schouten, R. M., Duivesteijn, W., Räsänen, P., Paul, J. M., and Pechenizkiy, M.** Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression. In *Proc. ECML PKDD* (2024), p. 66–82 [174].

8.1 Introduction

Learning mathematics is hard. At the neuro-cognitive foundation of young children's math development are core numerical processing competences such as the ability to enumerate small sets of dots and to compare the relative magnitudes between sets [59]. These numerical competences are diagnostic markers of emerging math abilities from as early as preschool age [71] which make them targets for conceptually motivated intervention programs [18].

We investigate characteristics that define exceptional patterns of young children's enumeration ability. Generally, enumeration performance reflects two distinct processes: the *subitizing* system where small sets (1-4 dots) are recognized accurately and rapidly, and the *counting* system where larger sets are enumerated more slowly perhaps by counting or other enumeration strategies [157]. Figure 8.1 gives an example; the enumeration response time of small sets is relatively flat while the counting slope is steeper. The inflection point demarcates the subitizing range from the counting range.

Individual differences in subitizing range predict math ability [130]. An inability to subitize is associated with dyscalculia [115]. There is value in accurately and reliably estimating the parameters that define subitizing patterns (initial reaction time, range, slope). Common algorithms used for estimating the subitizing range can produce inconsistent results [120], especially among individuals with dot enumeration curves that deviate from the typical curve.

We develop a piecewise linear regression model class for Exceptional Model Mining (EMM) [54] to discover subgroups of children whose subitizing curves exhibit atypical patterns. EMM is a local pattern mining framework seeking coherent subgroups in a dataset that somehow behave exceptionally. We develop various quality measures based on log likelihood that allow us to discover atypical subitizing patterns such as deviating initial reaction times, subitizing ranges, counting slopes, or a combination of those.

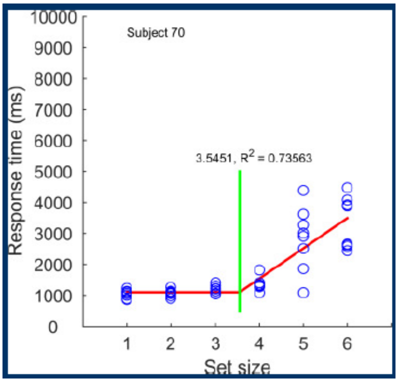
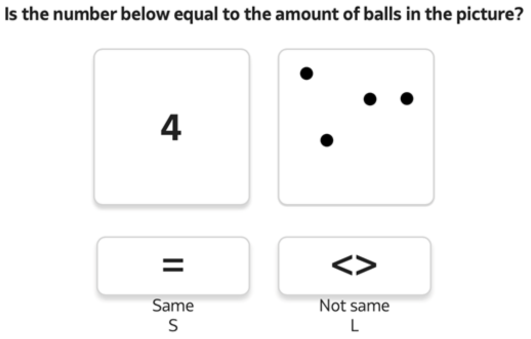


Figure 8.1: Dot enumeration response time regressed on set size (number of dots) using segmented linear regression with one break point.



Is the number below equal to the amount of balls in the picture?

Figure 8.2: An example of a dot enumeration item for set size 4; the correctly answered items are used in Figure 8.1.

We use data collected by the FUnctional Numerical Assessment (FUNA) study [155]. Numerical processing competences and math abilities are assessed using several computer-assisted *tasks*. Some of these tasks contain a fixed number of questions, or *items*; others are time-based and the number of answered items will vary per child. Items are taken from a larger set of items, and not necessarily answered in the same order. Consequently, the dataset does not follow the conventional data mining representation where each individual can be described with one tuple of attribute values, and where a column contains the same semantic information for each individual.[1] Hence, pre-processing is required to allow existing algorithms to search through the space of candidate subgroups. We discuss a manner tailored to the item-performance data at hand.

The main contributions of this chapter are: 1) an EMM model class and various quality measures for segmented linear regression; 2) a deeper understanding of how subitizing patterns relate to other numerical processing competences and emerging math abilities; 3) an effective pre-processing technique for handling repeatedly measured attributes in descriptive space.

8.2 The FUnctional Numerical Assessment study

The FUnctional Numerical Assessment (FUNA) project [155] is a large-scale research program in Finland to develop digital assessment tools for detecting dyscalculia and dyslexia. Currently, several studies are run to evaluate the validity and reliability evidence of the tasks [78]. The current version has been normed in Finnish and Finnish-Swedish languages for grade levels 3 to 9 (9 to 15 years old).

In the FUNA-DB (Dyscalculia Battery) the children respond to six digital (CAI) *tasks* using a tablet or a computer: Number Comparison (NC), Dot Matching equivalence task (DM), Single Digit Addition (SA), Single Digit Subtraction (SS), Combination Addition (CA) and Number Series (NS). Every task consists of multiple questions, or *items*. The tasks SA, SS, CA, and NS measure arithmetic fluency, and the items considered easier are provided earlier than more difficult items, but the exact order is not the same between children (i.e., quasi-random). In the number processing tasks (NC, DM), a set of predefined items are presented in a fully random order. Figure 8.2 displays an example of a DM item. Children compare a symbolic number (1-9) to a non-symbolic representation of a number. The location of the dots is randomized as well. When the symbolic and non-symbolic representations are the same, and when the children answer correctly, the DM task can be considered a Dot Enumeration (DE) task: determining the number of dots in a visual array.

Table 8.1 displays a dataset slice. On the right side (to be used as *target* attributes in the EMM model class, see Section 8.6), we present information from the DE task. Attributes ℓ_1 and ℓ_2 represent the set size (1-9) and response time in milliseconds respectively. These

[1] Children build up experience with the type of tasks at hand while the study unfolds. Suppose that two children perform Task T, but Child A is given this task earlier in the procedure than Child B. Then, Child B will have built up more experience than Child A with similar tasks, before executing Task T. A conventional data mining representation of this data would record performance of both children on Task T in the same column, but this belies the reality that these performances are not measured in an equal manner.

Table 8.1: Small slice of FUNA dataset. Some descriptors originate from the NC-task (a_2,a_3,a_4), others from the SA, SS, or CA task (not shown here), or from the general background information (sex, a_1). In our EMM instance, target attributes originate from the DE task (ℓ_1, ℓ_2). All task-based attributes contain data from multiple items, resulting in tuples of values. The number of values per tuple may vary per child and per task.

	sex	NC			DE	
i	a_1	a_2	a_3	a_4	ℓ_1	ℓ_2
1	f	(4,3,1,4,...)	(1,0,0,...)	(1200,1150,...)	(5,1,8,...)	(1330,14...)
2	f	(2,3,1,7,...)	(1,1,1,...)	(1240,1510,...)	(8,2,1,...)	(2630,21...)
3	m	(5,2,8,6,...)	(0,1,1,...)	(1490,1250,...)	(4,9,2,...)	(2130,19...)
4	f	(8,2,1,5,...)	(0,1,1,...)	(1180,1120,...)	(7,4,5,...)	(2610,16...)

attributes are the dependent and independent variables in a segmented linear regression model class as visualized in Figure 8.1.

We indicate the fact that we obtain data from multiple DE items per child, by using tuples (e.g., for the first item of child 1, the set size was 5 and response time was 1330 ms). For the SA, SS, and CA tasks, the number of items (tuple-length) differs per child; for the NC and DE tasks, the tuple-length is 52.

Apart from the set size and response time for each task, we may consider information such as whether the item is answered correctly, what is the correct answer, and what is the numerical distance between two numbers shown in a certain item. All this information is represented as separate attributes (e.g., attribute a_3 indicates where the items on the NC task have been answered correctly (1) or not (0)) and will be used to discover and *describe* exceptional subgroups of children. We also have some descriptive information, such as a child's sex (a_1), grade, and the language (Finnish or Swedish) in which they executed the tasks.

The data format as used by most traditional data mining algorithms is also known as a propositional table; these are single-table representations where each individual can be described with one term. In the attribute-value case, this term is a tuple of attribute values [116]. For instance, a student could be represented by a three-tuple specifying age, grade and language. Generally in EMM, we let the subgroup description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathscr{A}_j$ of the respective attribute a_j. For instance, a description $sex = girl \wedge language = Finnish$ covers all girls who executed the FUNA tasks in Finnish.

However, for all attributes other than sex, grade and language, our dataset does not follow this conventional data mining representation; a descriptive attribute is not associated to one value, but rather to a tuple of values. In this case, it is unclear what it means to apply a selector sel_j directly; a selector $a_2 \leq 3$ would select items rather than individuals and a selector such as $a_{2t} \leq 3$ where t refers to the item indicator, would inflate the number of descriptors, which is detrimental to efficient traversal of the search space. In addition, such a selector has little conceptual meaning, again because the items are quasi-randomly ordered and item t is not the same across children. We will provide a more satisfactory alternative in Section 8.5.

8.3 Background

Exceptional Model Mining (EMM) was introduced in Section 2.3. In short, EMM [54] is a Local Pattern Mining (LPM) framework seeking coherent subgroups in the dataset that somehow behave exceptionally. The attributes are divided into descriptors $a_1, \ldots, a_k$ and targets $\ell_1, \ldots, \ell_m$. Dataset Ω is then a bag of n records $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m). \tag{8.1}$$

Formal definitions for subgroup descriptions and quality measures were given in Definitions 2.1 and 2.2, respectively.

In traditional EMM, the combination of Equation (8.1) and a description language based on conjunctions of selection conditions implicitly assumes the data to be in a flat-table format where every record is an individual that is described by a tuple of attribute values, and placed on a new row in the single flat-table. In contrast, in this chapter, an attribute a or ℓ may or may not be measured repeatedly per individual i. We focus our notation on the descriptive attributes, and write a^i_{jt} to denote the t^{th} measurement of the j^{th} descriptive attribute for the i^{th} individual. Compared to Equation (2.1), the form of the descriptive part of individual $r \in \Omega$ changes to:

$$r = \big((a_{11}, a_{12}, ..., a_{1t}, ..., a_{1t_1}), (a_{21}, ..., a_{2t_2}), ..., (a_{k1}, ..., a_{kt_k})\big), \tag{8.2}$$

where t^i_j refers to the number of repeated measures of attribute a_j for individual $i \in \{1, 2, ..., n\}$, which may vary across individuals and attributes; we let $t_j = \max_{i=1,2,...,n} t^i_j$. Some descriptors may be measured only once per individual (such as sex in Table 8.1); then, $t^i_j = 1$ for all i.

8.3.1 Segmented linear regression

The goal of regression is to predict the value of an attribute y given a new value of $\mathbf{x}$, where $\mathbf{x}$ is a random draw from a vector of variables $\mathbf{X} = (X_1, \ldots, X_d)$. The simplest linear model for regression is one that involves a linear combination of the input variables and parameters $\mathbf{w}$: $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. We additionally aim to model the uncertainty, modeling a predictive distribution $p(y|\mathbf{x})$ by assuming that the deterministic function $f(\mathbf{x}, \mathbf{w})$ has additive Gaussian noise with zero mean and precision β (inverse variance). We then obtain the likelihood function:

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1}) = \prod_{i=1}^{n} \mathcal{N}(y[i] \mid \mathbf{w}^T\mathbf{x}[i], \beta^{-1}), \tag{8.3}$$

Next, estimating $\mathbf{w}$ and β using Maximum Likelihood Estimation shows that the log likelihood of a regression model depends on the sum-of-squares error function (SSR) [23] (see [173, Section 1] for an elaboration):

$$\ln p(y|\mathbf{x}, \mathbf{w}, \beta) \approx SSR(y, f(\mathbf{x}, \mathbf{w})) = \sum_{i=1}^{n} \big(f(\mathbf{x}[i], \mathbf{w}) - y[i]\big)^2.$$

Segmented linear regression appears to require non-standard optimization techniques. However, one can parameterize the model such that it can be modeled using an iterative, linear approach [141]. We focus on modeling two line segments between response variable y and one explanatory variable x_h by fitting the terms:

$$y = g(x_h, \alpha, \beta, \psi) = \alpha x_h + \beta(x_h - \psi)_+ \tag{8.4}$$

where $(x_h - \psi)_+ = (x_h - \psi) \cdot I(x_h > \psi)$ where $I(\cdot)$ is the indicator function equal to 1 if the statement is true and 0 otherwise. Consequently, ψ is the x-axis break point, α is the slope of the line segment to the left of ψ, and β is the difference in slopes between the line segments to the left and right of ψ. Next, [141] iteratively fit linear models of the form $\alpha x_h + \beta U^{(s)} + \gamma V^{(s)}$ with $U^{(s)} = (x_h - \psi^{(s)})_+$ and $V^{(s)} = -I(x_h > \psi^{(s)})$. Every iteration, $\hat{\psi}^{(s+1)}$ is updated through $\left(\psi^{(s+1)} - \psi^{(s)}\right) = \hat{\gamma}/\hat{\beta}$ and when the algorithm stops and $\hat{\gamma} \approx 0$, the s^{th} approximation is the Maximum Likelihood Estimate: $\hat{\psi}^{(s)} \equiv \hat{\psi}$ [141].

8.4 Connections to existing SD/EMM approaches

Linear target models for EMM are not a new concept [143]. Existing model classes use QMs comparing a regression parameter between the subgroup and a reference model. Instead, we follow the approach of [188] and [170] who build QMs on the log likelihood. These QMs do not directly compare parameter estimates but rather evaluate the overall fit of a model estimated on the subgroup. In addition, in this chapter, we utilize the special situation that when we assume Gaussian noise, maximizing the log likelihood is similar to minimizing the residuals sum-of-squares. This characteristic simplifies the notation and calculation of our QMs.

Our dataset has a nested structure: we aim to create subgroups at the level of the individual, while having access to repeated measures per individual in both target and descriptive space. We are not the first to consider time-varying target attributes. For instance, [170] analyze blood glucose fluctuations and [26] discover funding applications with deviating temporal sub processes. However, in descriptive space, these authors use attributes that are measured at the same level as the individual; their flattening approach can be categorized as a transformation to a wide flat-table data format. Alternatively, [124] transformed their data into a long, stacked flat-table format where each row contains a transition rather than an entire sequence. To this end, [124] had to change their notion of an individual. Under the hood, some form of *propositionalization* [116] takes place in [26, 124, 170], transforming hierarchical data with one-to-many relations into a single-table representation where each individual can be described with one term.

Relational subgroup discovery (RSD) [211] uses a proportionalization-based approach that accepts feature language declarations similar to those used in Progol [142]. Our proposed method is best described as a simple *aggregation* approach to feature construction [112]. We do not apply automated feature construction methods; these typically assume that columns of the dataset have a coherent semantic meaning, which our data does not (cf. Footnote 1). We show that with domain-specific aggregation functions, subgroup interpretability blossoms.

8.5 Our proposed flattening approach

An *aggregated descriptor* is a descriptive attribute constructed out of one or more original descriptors, where the original descriptors are defined as in Section 8.3 and may or may not contain repeated measures per individual. The goal is to describe each individual with one tuple of attribute-values as in Equation (8.1), rather than a tuple of tuples as in Equation (8.2). This allows defining descriptions as conjunctions of selection conditions over the aggregated descriptors.

Denoting an original descriptor with a_j, we construct an aggregated descriptor $\tilde{a}_h$ by applying a function $\xi : \mathbb{R}^* \to \mathbb{R}^1$ such that per individual, the number of observed values on attribute $\tilde{a}_h$ is 1. A function ξ may be applied to one or more time-varying descriptors, possibly in combination with an invariant descriptor.

Definition 8.1 (Aggregated descriptor)**.** *Given one or more descriptors $a_* \subseteq \{a_1, a_2, ..., a_k\}$, an aggregated descriptor $\tilde{a}_h$ is an attribute constructed by applying a function $\xi : \mathbb{R}^* \to \mathbb{R}^1$ such that per individual, the number of observed values on attribute $\tilde{a}_h$ is 1, i.e.: $\tilde{a}_h = \xi(a_*)$ with $a_* \subseteq \{a_1, a_2, ..., a_k\}$.*

Aggregated descriptors may arise from a function such as a summation or average, they may be non-linear (conditional) functions of one or more original descriptors, and/or they could be parameter estimates of a statistical model. Section 8.5.1 provides examples of all of these for the FUNA study.

The aggregated descriptors induce a tweak to the definition of a subgroup:

Definition 8.2 (Subgroup)**.** *A subgroup corresponding to description D is the bag of records $G_D \subseteq \Omega$ that D covers:*

$$G_D = \{r^i \in \Omega | D(\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_s) = 1\}. \tag{8.5}$$

The descriptive domain $\tilde{\mathscr{A}}$ is the collective domain of all aggregated descriptors $\tilde{a}_1, \ldots, \tilde{a}_s$ and the time-invariant descriptors $a_\dagger = \{a_j \in \{a_1, \ldots, a_k\} | t_j = 1\}$.

8.5.1 Domain-specific aggregations functions

Definition 8.1 allows for many variations. In the context of FUNA, a simple example is a function ξ_{max} that counts the number of answered items per task. For instance, $\tilde{a}^i_1 = \xi_{\text{max}}(a^i_{\text{NC}}) = t^i_{\text{NC}}$ is the number of NC items answered by individual i, where a_{NC} is the item-indicator of task NC. We may want to know how many items individual i answered correctly: $\tilde{a}^i_2 = \xi_{\text{sum}}(a^i_3) = \sum_t a^i_{3t}$, where a_3 is a binary attribute as in Table 8.1. We could subsequently measure the proportion of correctly answered NC items as follows: $\tilde{a}^i_3 = \xi_{\text{max}}(a^i_{\text{NC}})/\xi_{\text{sum}}(a^i_3)$.

Other aggregation functions that are interesting from a domain perspective are the mean and median response time of the correctly answered items. We write $\tilde{a}^i_4 = \xi_{\text{meanTC}} = (\xi_{\text{sum}}(a^i_3))^{-1} \cdot \sum_{t \in \{1,...,t_4\} \text{ s.t. } a^i_{3t}=1} a^i_{4t}$. For ξ_{medianTC} we would do something similar but take the median rather than the mean.

In the domain of educational learning, the Inverse Efficiency Score (IES) [71] is a measure that combines both the median response time and the accuracy (proportion of correctly

Table 8.2: An overview of the aggregation functions used in FUNA.

Tasks	Name	Explanation
NC,SA,SS,CA	MaxItem	Number of answered items
NC,SA,SS,CA	SumAnsC	Number of correctly answered items
NC,SA,SS,CA	PropAnsC	Proportion of correctly answered items
NC,SA,SS,CA	MeanTC	Mean response time of correctly answered items
NC,SA,SS,CA	MedTC	Median response time of correctly answered items
NC,SA,SS,CA	IES	Inverse Efficiency Score
NC	IcNumD	Intercept of the response time regressed on the distance between the two numbers of every item
NC	SlNumD	Slope of the response time regressed on the distance between the two numbers of every item
NC	IcNumR	Intercept of the response time regressed on the ratio between the distance and the largest of the two numbers of every item
NC	SlNumR	Slope of the response time regressed on the ratio between the distance and the largest of the two numbers of every item

answered items). The IES allows researchers to identify children with high response times, or a low proportion of correctly answered items, since the IES score is high in both cases. For an individual:

$$\tilde{a}^i_6 = \xi_{\text{IES}}(a^i_{\text{NC}}, a^i_3, a^i_4) = \frac{\xi_{\text{MedianT}}(a^i_4)}{\xi_{\text{PropAnsC}}(a^i_{\text{NC}}, a^i_3)} = \frac{1/t^i_4 \sum_t a^i_{4t}}{t^i_{\text{NC}} / \sum_t a^i_{3t}}. \tag{8.6}$$

For the Number Comparison (NC) task, it is interesting to analyze the numerical distance effect [83]. When tasked with saying which of two numbers is greater, this task is easier to perform when the numbers are far apart (*NumD*). If numbers have the same distance, the task is hypothesized [157] to be easier if the largest number is smaller. This is called the Number Ratio (*NumR*). We regress the response time of the NC items on the NumD (and once more for NumR), and evaluate the intercept (Ic) and Slope (Sl) of these models. Thus, we first create a time-variant descriptor $a_{\text{NumD}} = |a_{\text{NCL}} - a_{\text{NCR}}|$ (where a_{NCL} and a_{NCR} are the numbers shown on the left and right in each NC item) and then fit a linear regression model per individual i: $a^i_4 = f(a^i_{\text{NumD}}, w^i_0, w^i_1)$. Parameter estimates w^i_0 and w^i_1 are the intercept and slope of the regression model, stored as aggregated descriptors $\tilde{a}^i_7 = w^i_0$ and $\tilde{a}^i_8 = w^i_1$. We take the same approach for NumR.

An overview of these aggregated descriptors is given in Table 8.2.

8.6 Our proposed target model

We seek subgroups of children with atypical dot enumeration curves. We use the segmented linear regression model as a target model (cf. Section 8.3.1) with response time ℓ_2 as output (y) and set size ℓ_1 as input (x_h) (cf. Table 8.1). We are interested in finding any kind of deviation from the typical DE curve; in a typical DE curve the subitizing slope is close to zero, the subitizing range is somewhere between 3 and 4, and the counting slope is relatively steep.

Following [188] and [170], we assume that the parameters of a linear model fitted on the subgroup will likely describe the subgroup better than the parameters estimated on the entire dataset. Then, in the presence of a subgroup, the log likelihood of dataset Ω will increase if the parameters of the subgroup are separately estimated. For any subgroup SG and its complement SG^C,

$$\ln p(SG|\theta^{SG}) + \ln p(SG^C|\theta^{\Omega}) > \ln p(SG|\theta^{\Omega}) + \ln p(SG^C|\theta^{\Omega}),$$

where $\ln p(SG|\theta^{SG})$ is the log likelihood of the subgroup for a segmented linear regression model estimated on the SG with $\theta^{SG} = (\alpha^{SG}, \beta^{SG}, \psi^{SG})$. We expect this term to be larger than the log likelihood of the subgroup for a segmented linear regression model estimated on the entire dataset Ω: $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$. Next, we use the characteristic of linear regression that maximizing the log likelihood is similar as minimizing the sum-of-squares error function (SSR) (see Section 8.3.1, and [173, Section 1]) and aim to find subgroups where $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$ holds. Hence, we formulate our first QM as follows:

$$\begin{aligned} \varphi_{\text{ssr}} &= \frac{1}{\varphi_{\text{ef}}} \cdot - \frac{A}{N^{SG}} \\ A = SSR(\ell_2, g(\ell_1, \theta^{SG})) &= \sum_{i=1}^{n^{SG}} \sum_{t=1}^{t^i_{\ell_1}} \left(\ell^i_{2t} - \hat{\alpha}^{SG} \ell^i_{1t} - \hat{\beta}^{SG} (\ell^i_{1t} - \hat{\psi}^{SG})_+ \right)^2, \end{aligned} \tag{8.7}$$

where $N^{SG} = \sum_{i=1}^{n^{SG}} t^i_{\ell_1}$ is the number of observations in the subgroup in target space and φ_{ef} is the entropy function [54] to discourage tiny subgroups. We take the SSR of ℓ_2 with respect to $g(\ell_1, \theta^{SG})$, which is defined in Equation (8.4). If the sum-of-squared error decreases, φ_{ssr} increases.

Although both the regression parameters and precision depend on the sum-of-squares, they are statistically independent. This means that we could find subgroups with a small error where $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$ does not hold; the log likelihood of the subgroup may be large, but it may not be larger than the log likelihood of the global model, for instance when the regression parameters θ^{SG} do not differ much from θ^{Ω}. Therefore, we propose a QM that rewards not only small values of SSR for the subgroup, but also values of SSR for the subgroup that are smaller than the SSR of the subgroup evaluated on the global model:

$$\varphi_{\text{ssrb}} = \varphi_{\text{ef}} \cdot \frac{A(B - A)}{N^{SG}}, \tag{8.8}$$

where A is as in Equation (8.7) and B is similar but with θ^{SG} replaced by θ^{Ω}.

8.7 Experiments on real-world data

We perform two experiments.[2] First, we randomly sample 5% of the children and experiment with both QMs φ_{ssr} and φ_{ssrb}. We perform beam search [54, Algorithm 1] with $b = 4$, $w = 20$, and $q = 10$. Especially when working with domain-specific data, we aim for our resulting subgroup set to be a good balance between interpretability, variety, and quality. To further understand how a weighted coverage scheme (WCS) [201] can contribute to finding such a balanced subgroup set, and what its relation is to the search depth d, we vary $d \in \{3, 5\}$ and the multiplicative weighting parameter of the WCS $\gamma \in \{0.1, 0.5, 0.9\}$. We evaluate our results by inspecting the average quality of the subgroup set, the average size of the subgroups, the number of subgroups (out of $q = 10$) that validation with the Distribution of False Discoveries (DFD) [55] cannot distinguish from false discoveries over $m = 50$, the average run time, and two measures of subgroup set redundancy: Joint Entropy (JE) [201] and median Jaccard similarity (JSIM) [153] (see [173, Section 2] for precise definitions). We use the `pwlf` Python library to fit our segmented linear regression models [92].

Second, based on our findings in the first experiment, we choose the most appropriate QM, value for d and value for γ, and repeat the experiment with the full FUNA dataset ($n = 15486$). Beam search width $w = 20$, $b = 4$ and $q = 20$. All these children have at least 5% of their answers correct in each descriptor task (NC, SA, SS, CA) and the children have at least one observed answer for every possible set size in the DE task. The maximum number of observed items in the DE task is 18 per child.

Extra experiments on Curran dataset

We perform an additional set of experiments on a fully public dataset and discover subgroups of children with exceptional relations between age and reading skills. Since our quality measures generalize to linear regression problems other than segmented linear regression, we perform these extra experiments with polynomial regression. More information and a short discussion of the results can be found in [173, Section 3].

8.7.1 Subgroup set redundancy and weighted coverage schemes

Figure 8.3 presents the standardized, average quality of a subgroup set ($q = 10$) for various values of d, γ, and both QMs. In essence, the results are as expected: the quality increases with the description length d and the weight parameter γ increases, and the impact of varying γ is larger for smaller d (see Figure 8.3; absolute difference between the smallest and largest quality for varying γ is larger for $d = 3$ than for $d = 5$).

Table 8.3 reports the other evaluation metrics: the average subgroup size decreases when either d or γ increase, and in general, the subgroup set redundancy is larger when d decreases or γ increases (higher JE, lower JSIM). Except for 2 subgroups for φ_{ssrb} when $d = 3$ and $\gamma = 0.1$, all discovered subgroups can be considered valid discoveries.

[2] Our experimental code, all results, and a slice of the FUNA dataset are available at `https://github.com/RianneSchouten/FUNA_EMM/`.

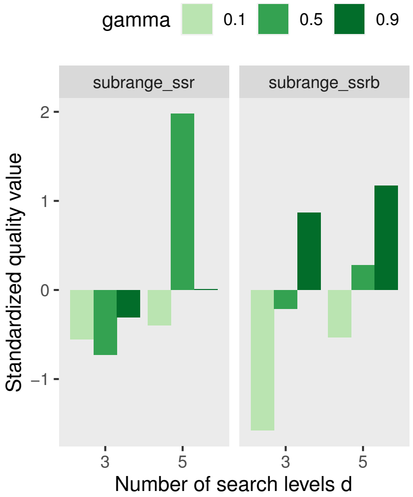


Figure 8.3: The relation between the average quality of a subgroup set ($q = 10$, standardized per QM), search depth d, and WCS parameter γ, for both QMs.

For φ_{ssr}, given d, the average subgroup size, JE, and JSIM are comparable when varying values of γ. It seems that there is barely an effect of the WCS. When $d = 5$, the average quality is lower for $\gamma = 0.9$ then for $\gamma = 0.5$, and when $d = 3$, the average quality is lower for $\gamma = 0.5$ than for $\gamma = 0.1$. These results are unexpected since a decreasing γ is supposed to increase the variety of the subgroup set at the cost of average quality. Inspecting the individual descriptions and qualities, we find that for φ_{ssr} the variety in the subgroup set is larger when $\gamma = 0.9$ than when $\gamma \in \{0.1, 0.5\}$. Most likely, the reason is the use of a *square* when calculating the quality. Even when we use a strict WCS, the same subgroup recurs, since the weighted quality of the other subgroups does not beat the non-weighted quality of the recurring subgroup. When the WCS is very strict (small γ), at lower search levels, important precursors may be removed and not available for refinement at higher levels. As a consequence, a subgroup set with a strict WCS could have fewer candidate subgroups, which in the end creates a relatively redundant subgroup set. It is unfortunate that JE and JSIM do not fully reveal these conclusions.

With φ_{ssrb} the subgroup sets are less redundant than with φ_{ssr}, especially for small values of γ. Clearly, JSIM increases and JE decreases when γ increases. Subgroups found with $d = 5$ are slightly smaller than for $d = 3$.

8.7.2 Exceptional learning behavior

We perform the experiment on the entire dataset with φ_{ssrb}, since this QM turns out to be stable and produces small and interesting subgroups. We choose $\gamma = 0.5$ to balance between high quality and low redundancy. We choose $d = 3$ since Table 8.3 shows that

Table 8.3: Experimental results for both QMs, $d \in \{3,5\}$, $\gamma \in \{0.1, 0.5, 0.9\}$.

QM	d	γ	Prop	DFD	JE	JSIM	Time
φ_{ssr}	3	0.1	0.20	0	1.36	0.87	1.63
		0.5	0.16	0	1.68	0.75	1.61
		0.9	0.16	0	1.55	0.73	1.62
	5	0.1	0.12	0	0.91	0.88	2.97
		0.5	0.13	0	0.79	0.91	2.95
		0.9	0.13	0	0.77	0.90	2.93
φ_{ssrb}	3	0.1	0.22	2	4.35	0.18	1.60
		0.5	0.08	0	2.35	0.31	1.56
		0.9	0.05	0	1.14	0.44	1.35
	5	0.1	0.06	0	2.19	0.29	2.17
		0.5	0.06	0	2.01	0.31	2.13
		0.9	0.05	0	1.08	0.46	1.86

these results do not differ much from $d = 5$, and a description with fewer literals is easier to interpret for domain experts. Descriptions and target models of all top-20 exceptional subgroups can be found in [173, Section 2]; we report a smaller selection in Table 8.4 and Figure 8.4.

Although we allow for descriptions to have $d = 3$ literals, strong performance is found in single-attribute subgroups. There is a variety in used descriptors (multiple aggregation functions, multiple tasks), subgroup size, and target models. Compared to the segmented linear regression parameters of the global model, 15 out of 20 exceptional subgroups have a subitizing range lower than average; the other 5 have a higher subitizing range.

Subgroups 1 and 2 have very similar subitizing curves: children in these subgroups are particularly slow to subitize, and these groups are the only ones that have an intercept over 2 seconds. The subgroups contain children with slow NC response times (either expressed in terms of IES or mean response time) and both are slow to solve addition problems (based on SA and CA tasks). The groups are small, and probably most typical of dyscalculia, or at the very least groups that are made up of children who are likely to have maths learning difficulties. The dyscalculia prevalence estimate is 3-6% [181], which is in accordance with the subgroup sizes 0.05 and 0.06 for subgroups 1 and 2 respectively.

Subgroup 5 is a more general version of subgroup 1; it covers 50% of the children and contains only the first literal. The subitizing curve shows the same trend as the one of subgroup 1, but less extreme: the subitizing range is smaller than the global model, but not as small as in subgroup 1, and intercept, subitizing slope, and counting slope are larger than in the global model, but not as large as in subgroup 1. Domain experts suspect that this subgroup may reflect maths learning difficulties as well.

Subgroup 6 is the inverse of subgroup 5. This is not only clear from the description in Table 8.4, but from the regression model in Figure 8.4 as well; the subitizing range is higher, and the intercept and subitizing slope are lower than in the global model. Subgroups 13, 15, 18, and 19 are the other four subgroups that have subitizing ranges above the average, and characteristically have subitizing intercepts (baseline response time or speed of

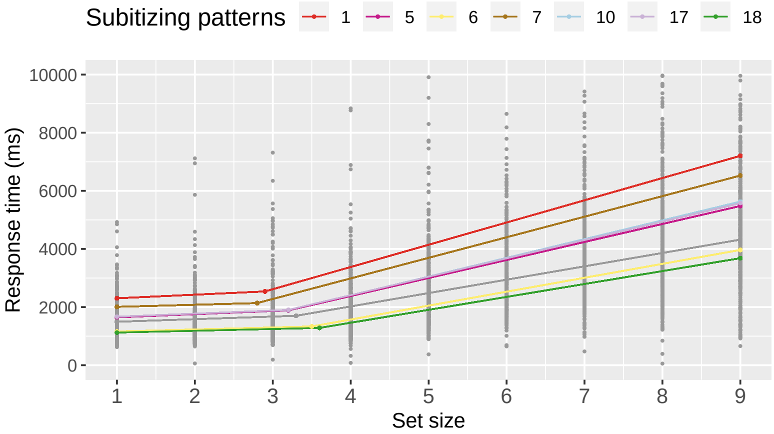


Figure 8.4: Estimated segmented linear regression models of subgroups 1, 5, 6, 7, 10, 17 and 18 discovered with φ_{ssrb}. Target model equations can be found in Table 8.4.

Table 8.4: Subgroup proportion, description and estimated target models for subgroups 1, 5, 6, 7, 10, 17 and 18, discovered with φ_{ssrb}. The global target model is $1407+88\ell_1+463(\ell_1-3.3)_+$.

SG	Prop	Description	Target model
1	0.05	NC-IES:(0.04,1.0) ∧ NC-MeanTC:(0.23,0.74) ∧ SA-MeanTC:(0.71,1.0)	$2179+124\ell_1+764(\ell_1-2.9)_+$
5	0.50	NC-IES:(0.03,1.0)	$1541+106\ell_1+624(\ell_1-3.2)_+$
6	0.50	NC-IES:(0,0.03)	$1091+\ \ 70\ell_1+475(\ell_1-3.5)_+$
7	0.12	NC-MeanTC:(0.16,1.0) ∧ SS-SumAnsC:(0.0,0.32)	$1938+\ \ 70\ell_1+712(\ell_1-2.8)_+$
10	0.38	SA-MeanTC:(0.71,1.0)	$1544+108\ell_1+641(\ell_1-3.2)_+$
17	0.30	grade = 3	$1561+106\ell_1+634(\ell_1-3.2)_+$
18	0.27	SA-MaxItem:(0.6,1.0)	$1064+\ \ 62\ell_1+441(\ell_1-3.6)_+$

processing) that are 300-350ms faster than the average and at least 500ms faster than any other group in the table. They also have shallower (faster) counting slopes by 150-200ms than most other groups.

Subgroups 18 and 19 have target models that are very similar to the one of subgroup 6, even though the descriptions of these subgroups differ. Subgroup 6 expresses the subgroup in terms of NC-IES while subgroup 18 does this in terms of an arithmetic addition task (SA). A similar things occurs for subgroups 5 and 10: the target models are similar while the descriptions use aggregated descriptors from different tasks. These findings suggests relations between number processing skills and arithmetic skills. They additionally show that it may be possible to obtain diagnostic information by focusing on fewer tasks; it may be possible to know the results on a particular task given the performance on another task. This is a promising result that provides opportunity for further development of assessment tools and intervention programs.

The only subgroup that does not use an aggregated descriptor is subgroup 17, which selects children in the third grade. Interestingly, the estimated target model of subgroup 17 is similar as the ones for subgroups 5 and 10; similar to the global model, expect for a larger counting slope. Compared to the other children in the FUNA dataset, the children in subgroup 17 are younger and hence, slower for all tasks, including the NC (subgroup 5) and SA (subgroup 10) tasks.

8.8 Conclusion

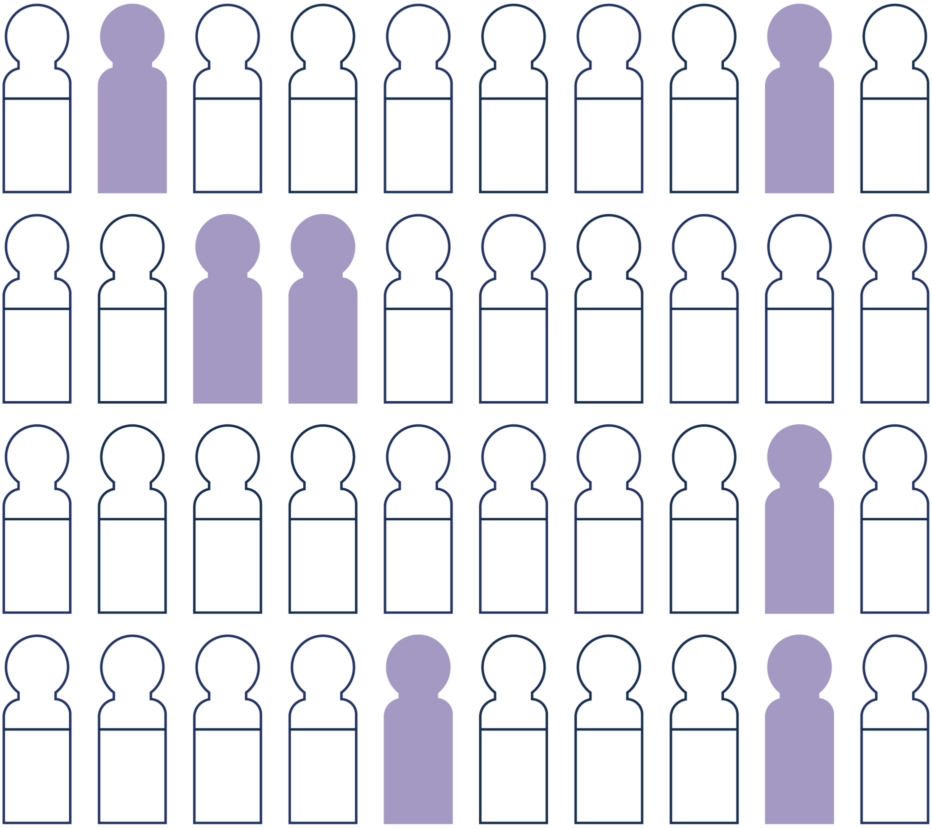
The FUnctional Numerical Assessment (FUNA) project [155] develops digital assessment tools for detecting dyscalculia and dyslexia in young children by evaluating numerical processing competences such as the ability to enumerate small sets of dots and to compare the relative magnitudes between sets. These numerical processing competences are diagnostic markers of children's emerging math abilities [71]. In this chapter, we particularly focus on the characteristics that define children's enumeration ability, such as the threshold at which children can determine the correct number of dots at a glance, known as subitizing range, and other parameters of subitizing patterns such as the initial reaction time and counting slope. Common algorithms used for estimating subitizing range can produce inconsistent results [120] especially among individuals with dot enumeration curves that deviate from the typical curve.

Therefore, we develop an EMM model class for segmented linear regression to discover subgroups of children whose subitizing curves exhibit atypical patterns. It could be argued that choosing segmented linear regression as a model class is a drawback since the observations are not independently distributed (i.e., a model is estimated on n^{SG} independent children, who all contribute the measurements of several items, resulting in a total number of N^{SG} observations). Despite of that, we follow this approach since segmented linear regression fits the neuro-cognitive concept of subitizing very well. Furthermore, the assumption of independent observations is required for most of the other algorithms as well; segmented linear regression has the least baggage built into it.

Our findings confirm the belief that numerical processing competences strongly correlate with arithmetic skills. We find several exceptional subgroups that confirm existing knowledge, including subgroups that are considered typical of dyscalculia; these children have slow NC response times and are slow to solve addition problems. We find subgroups with similar subitizing patterns but different descriptions. This indicates the strong relation between subitizing, counting, and arithmetic ability, and additionally provides promising opportunities for further development of assessment tools and intervention programs that focus on fewer tasks or a reduced number of items per task: it may become possible to know the results on a particular task given a child's performance on another task.

Both quality measures in this chapter assume that the overall population and subgroups are best modeled with the canonical subitizing range model: a piecewise linear regression model with precisely one break point. However, it is entirely possible that coherent subgroups of children do not follow this regimen: some groups may display no substantial break point; behavior of others might be best modeled by multiple break points. The piecewise linear regression model class for EMM can accommodate this sort of behavior, but it requires development of a new QM: log likelihoods will necessarily increase when more break points are available to the model, so some penalty for model complexity must be involved.

9



Conclusions

In the previous chapters of this dissertation, we presented new insights that contribute to answering the research question introduced in ***Chapter 1****: "How to discover exceptional subgroups in hierarchical data?" This chapter summarizes the main contributions of our work by highlighting where in our unified terminology the respective chapters of this dissertation fit in. We furthermore give an overview of domain-specific contributions and discuss directions for future research.*

9.1 Summary of contributions

In this dissertation, we analyze variation in human behavior using a Local Pattern Mining (LPM) framework called Exceptional Model Mining (EMM) [54, 121]. EMM aims to extract practically relevant patterns from data. However, real-world data often has a hierarchical structure where the observations on entities of one entity type are nested in the entities of another entity type. The concept stems from the idea that individual persons are influenced by the social groups or contexts to which they belong (and vice versa) [84, 125, 187]. The individuals and social groups are conceptualized as a hierarchical system of individuals nested in groups, and groups nested in larger groups. Such a hierarchical system can be generalized to the scenario where information is repeatedly measured per individual, resulting in data where observations are nested in individuals [84, 125, 187].

In hierarchical data, a shared context introduces a correlation structure between individuals belonging to that context, or between measurements belonging to the same individual. Then, a common data mining assumption that observations are independent is violated. Consequently, hierarchical data structures pose problems for the framework of EMM regarding whether and how hierarchical data can be formatted as a flat table, how selection conditions can be used to cover a group of individuals and how we can assess exceptionality. Therefore, the main research question in this dissertation is:

How to discover exceptional subgroups in hierarchical data?

To answer this question, in **Chapter 3** we first formally defined hierarchical data as a collection of measurements taken from various types of entities, where the measurements on entities of one entity type are nested in the entities of another entity type. We then proposed a unified terminology that classifies existing EMM methodologies based on whether descriptors and targets reside at lower, the same or higher hierarchical levels than the subgroup level, the hierarchical level of the entity type for which subgroups should be formed.

The work presented in this dissertation significantly contributes to the body of scientific literature on EMM for hierarchical data. In Table 9.1, we reproduce the unified terminology presented in Table 3.1 and additionally highlight in blue where the respective chapters of this dissertation fit in. In sum, our work further populates **Box D** (sequential data in target space) and is one of the first to consider hierarchical data at a higher level than the subgroup level (**Box F**) and to consider hierarchical data with nested observations in both descriptive and target space (**Box A**). More specifically, the contributions of this dissertation are as follows:

1. We developed new EMM methodology for sequential data in target space (entity type *time* nested in *individuals*). We answered **EMM-RQs A1, A2 and A3** by utilizing Markov chains of varying order and by developing three new quality measures based on information-theoretic scoring functions. These quality measures allow for the number of parameters to differ between subgroup and population. They also naturally handle varying sequence-lengths (**Chapters 4 and 5**).

   Our proposed method further populates **Box D** in Table 9.1 ([170]). In contrast to [124], who introduced an EMM model class for 1st order Markov chains and worked

Table 9.1: Our proposed unified terminology for EMM for hierarchical data from Table 3.1. Descriptors and targets can be measured at lower, the same or higher hierarchical levels than the subgroup level. Here, we highlight the contributions of this dissertation in blue.

		Targets		
		lower	***same***	***higher***
Descriptors	***lower***	**A** [88, 93, 129] [174]	**B** [66, 77, 103, 114, 209, 211] [57, 132] [131, 138, 159]	**C**
	same	**D** [42, 72, 156, 204] [26, 143] [170] D^* [9, 20, 21, 47, 98] D^* [95]	**E** [17, 51, 54, 56, 109, 121] E_1^* [10, 50, 124] E_2^* [29, 85, 86, 106]	**F** [158] [171] F^* [19]
	higher	**G** G^* [185]	**H** [118, 119, 140] [16, 147, 202, 203]	**I**

around the hierarchical data structure by adopting a long flat-table data format (*Box E**), we let the sequential structure in target space intact.

Within Box D, our work is close to [26], who analyze Dynamic Bayesian Networks (DBNs) in target space and define a mismatch score between the subgroup and its complement. Instead, we compared the subgroup to the entire dataset because this is conceptually more relevant and computationally more efficient.

2. We developed EMM for Repeated Cross-Sectional data (EMM-RCS) (entity type *individual* nested in *measurement occasions*). Our proposed methodology answers **EMM-RQs B1, B2 and B3** by handling practical RCS data problems, including uneven spacing of measurements over time, fluctuating sample sizes and missing data. We developed a generic quality measure that provably discovers exceptional trend behavior by balancing measures of the distance to the population model, the uncertainty of those measures and the subgroup size in determining the exceptionality of a subgroup (**Chapters 6 and 7**).

   We are the first to propose a model class (exceptional trend behavior) that resides at a higher hierarchical level than the subgroup level (individuals). Consequently, our work is one of the first to populate **Box F** in Table 9.1 ([171]).

   Two other EMM-related approaches were positioned in Box F, but those methods consider a scenario where the *target attributes* form a hierarchy [19, 158]. Then, data values at higher hierarchical levels can be calculated directly from the lower level measurements; there is no uncertainty of measurements in hierarchical levels other than those at the lowest level. In contrast, throughout this dissertation, we consider hierarchical data where *entity types* form a hierarchy. Then, each entity type has its own associated set of attributes and measurement uncertainty occurs at every hierarchical level. This effect is particularly prevalent in RCS data where varying measurement occasions create that data distributions change over time.

3. We developed an EMM model class for nested data in target space and simultaneously solved problems regarding handling nested data in descriptive space (*items* nested in *digital assessment tasks*).

   Specifically, we answered **EMM-RQ C3** by developing a piecewise linear regression model class with various quality measures discovering subgroups of children whose subitizing curves exhibit atypical patterns. Furthermore, we answered **EMM-RQ C1 and C2** by 1) investigating existing propositionalization approaches (most of them are in Box B) and 2) proposing the concept of aggregated descriptors as a generic approach to flattening nested data in descriptive space (**Chapter 8**).

   We are one of the first to develop an EMM methodology for hierarchical data with nested observations in both descriptive and target space. Consequently, our work is one of the first to populate **Box A** in Table 9.1 ([174]).

   Within Box A, our approach significantly differs from [93], who cut time series into slices and construct features per slice. Then, the subgroup level changes from independently sampled time series to slices. In contrast, our features describe entire sequences. Furthermore, [88, 129] analyze sequences of items in descriptive space and extract two types of patterns: sets of items and subsequences. They focus on understanding (the order of) items per individual. In contrast, our individuals are described by multiple sequences (of binary, numerical and categorical type) and the interest is in extracting groups of individuals that share common traits. In target space, [88, 129] average top-N recommendations into one numerical value per user (SD) whereas we develop a target model that contains all nested values (EMM).

Throughout this dissertation, we demonstrated *internal* validity of our proposed methods by means of synthetic data studies and experiments on public, real-world data. In addition, we performed experiments to evaluate certain properties of our search strategy: two new refinement strategies for handling incomplete descriptors (**Chapter 6**), the effectiveness of (combining) several solutions for reducing subgroup set redundancy and validating the discovered subgroups (**Chapter 7**) and the interaction effect of applying a Weighted Coverage Scheme (WCS) [117, 201] and beam search parameter d (**Chapter 8**).

We uncover the effect that a more strict weighting regime (smaller value for γ in Equation 2.3) yields a more diverse subgroup set. In other words, it may occur that at early search levels important precursors are pushed out of the beam, rendering them unavailable for refinement at subsequent levels. A similar effect occurred when handling incomplete descriptors: increasing the space of candidate subgroup descriptions reduced the quality of the discovered subgroups because important precursors were removed.

9.2 Domain-specific contributions

Moreover, we demonstrated *external* validity of our proposed methodologies by enabling domain experts to confirm existing hypotheses and to spawn interest for new theories. We show that our proposed EMM methodologies are hypothesis-generating sources that, due to their exploratory nature, work as a starting point in further understanding variation in human behavior.

Diabetes care

Domain experts and clinicians hypothesized the use of iCGM-derived parameters for establishing individualized glycemic treatment [38, 46] (**D-RQ Diabetes care**). In this dissertation, we successfully deployed our proposed methodology and provided supporting evidence for this hypothesis. Specifically, we discover a variety of subgroups with exceptional blood glucose fluctuations: some subgroups transition towards higher blood glucose values; others towards similar or lower values. For instance, patients with high HbA_{1c} values are more likely to have a Time Above Range (TAR) that is too high. If those patients are also older than average, they are additionally less likely to have a good Time In Range (TIR). In contrast, patients with low HbA_{1c} values are likely to transition away from high blood glucose levels. Domain experts confirmed our findings (**Chapter 5**).

Public health

We contributed to answering **D-RQ Public health** by successfully deploying EMM-RCS to discover (combinations of) social group memberships that display exceptional deviations in the trend in adolescent alcohol use in the Netherlands. Using our generic quality measure, we discovered three types of exceptional trend deviations. Domain experts interpreted our findings and concluded that they confirm existing knowledge that age, educational level, and migration background are important descriptors of monthly alcohol use and that an interplay effect exists with life satisfaction, urbanization degree, and truancy. Furthermore, our findings spawned two hypotheses for further sociological research and provided dis-confirming evidence to a sociological hypothesis that aligns with existing studies (**Chapter 7**).

Learning analytics

We contributed to the understanding of individual variation in learning behavior. Specifically, we contributed to answering **D-RQ Learning analytics** by deploying our proposed EMM model class for segmented linear regression and confirmed the belief that numerical processing competences strongly correlate with arithmetic skills. Together with domain experts, we discovered several exceptional subgroups, including subgroups that are considered typical of dyscalculia; these children have slow Number Comparison (NC) response times and take longer to solve addition problems. Furthermore, we discovered subgroups with similar subitizing patterns but different descriptions. According to domain experts, these findings reflect the strong relation between subitizing, counting, and arithmetic ability. In addition, our findings reveal promising opportunities for further development of assessment tools and intervention programs that focus on fewer tasks or a reduced number of items per task. It may become possible to predict the results on a particular task given a child's performance on another task (**Chapter 8**).

9.3 Future work

During our work on the contributions in this dissertation, we identified many valuable research directions. In Section 3.6, we gave a short overview of research gaps regarding EMM for hierarchical data. There, we suggested to increasingly utilize ontologies to traverse the search lattice efficiently [118], to consider target models specifically designed for analyzing hierarchical data, such as location-scale models [75] and to focus on improving the process of aggregating lower-level measurements in descriptive space [198].

Furthermore, our unified terminology in Table 9.1 (cf. Table 3.1) reveals that there exist only few EMM methodologies for data with attributes that reside at a higher hierarchical level than the subgroup level (Boxes C, F, G, and I). Especially Boxes A, C, G and I are sparsely populated or even completely empty. In these corner boxes, there is space for further developing the EMM framework towards hierarchical data where attributes reside at a non-subgroup level, in descriptive and in target space.

Remark that hierarchical data is a form of non-IID data. Therefore, our classification covers all relevant work on SD and EMM for non-IID data. Yet, so far, most existing work on EMM for non-IID data considers hierarchical data. EMM instances developed for non-IID data that cannot be considered hierarchical according to Definition 3.2, are assigned to boxes with an asterisk. For instance, exceptional subgraph mining is positioned in *Box D**, since we can consider attributed vertices [9, 20, 21, 47] or attributed edges [98] to reside at a lower hierarchical level than the subgraph (the subgroup level). As a next step, it would be interesting to explore whether EMM methodology could deal with graphs that contain both attributed vertices and attributed edges, and to investigate graph data that contains observations on entities of multiple entity types. Other examples of non-IID data worth exploring with EMM are multi-relational databases with many-to-many relationships [209, 210], multivariate time-series data [85, 93] and unstructured data [6].

In addition, we observe a trend that domain experts increasingly use data collection methods other than the traditional research designs. Consequently, we expect a need for developing the EMM framework towards data with multiple modalities. For instance, in DIALECT-2 [67], besides measuring blood glucose values with an iCGM device, sensors collect repeated measurements of heart rate and step count, questionnaires give a subjective indication of patients' health behavior and food diaries provide information on calorie intake [67]. It would be valuable to develop EMM methodologies that take into account all this information. An example of an EMM instance developed for multi-modal data is [50], who discover exceptional spatio-temporal behavior.

Given the societal impact of the patterns discovered with our developed methodologies, we believe EMM has great potential to be integrated into real-world applications and software. An example of other recent work that demonstrates this potential future direction is [88], who first, develop an SD model class for extracting explanations of top-N recommendations made by a state-of-the-art recommender system and second, construct a global model, based on the set of explanations generated for n users, that acts as a recommender system itself. Alternatively, at a smaller scale, we see potential in deploying EMM at an individual level, solely using person-specific data. For instance, an app with an EMM implementation could have a signaling function by discovering exceptional deviations from an individual's typical behavior pattern.

To realize the implementation of EMM instances into real-world applications, the EMM framework should consider moving towards extracting patterns from data streams rather than fixed-size datasets, and towards discovering sequential patterns (almost) in real time.

Such advancements include developing EMM methodologies for understanding model change [7, 144] and concept drift [81]. Remark that non-stationary data streams where variable distributions change over time are closely related to Repeated Cross-Sectional data where independent observations are nested in measurement occasions. Consequently, the work presented in this dissertation could serve as a starting point in further exploring the possibilities of implementing EMM methodologies in real-world applications.

References

[1] **Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I., et al.** Fast discovery of association rules. *Advances in Knowledge Discovery and Data Mining 12,* 1 (1996), 307–328.

[2] **Agrawal, R., Srikant, R., et al.** Fast algorithms for mining association rules. In *Proc. VLDB* (1994), vol. 1215, pp. 487–499.

[3] **Akaike, H.** Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike,* E. Parzen, K. Tanabe, and G. Kitagawa, Eds. Springer, 1998, pp. 199–213.

[4] **Akaike, H.** A new look at the statistical model identification. In *Selected papers of Hirotugu Akaike,* E. Parzen, K. Tanabe, and G. Kitagawa, Eds. Springer, 1998, pp. 215–221.

[5] **Anscombe, F. J.** Graphs in statistical analysis. *The American Statistician 27,* 1 (1973), 17–21.

[6] **Arab, A., Arora, D., Lu, J., and Ester, M.** Subgroup discovery in unstructured data, 2022. `https://arxiv.org/abs/2207.07781`.

[7] **Artelt, A., Hinder, F., Vaquet, V., Feldhans, R., and Hammer, B.** Contrasting explanations for understanding and regularizing model adaptations. *Neural Processing Letters 55,* 5 (2023), 5273–5297.

[8] **Atzmueller, M.** Subgroup Discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 5,* 1 (2015), 35–49.

[9] **Atzmueller, M., Doerfel, S., and Mitzlaff, F.** Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences 329* (2016), 965–984.

[10] **Atzmueller, M., Mueller, J., and Becker, M.** Exploratory subgroup analytics on ubiquitous data. In *International Workshop on Mining Ubiquitous and Social Environments* (2013), pp. 1–20.

[11] **Battelino, T., Danne, T., Bergenstal, R. M., Amiel, S. A., Beck, R., Biester, T., Bosi, E., Buckingham, B. A., Cefalu, W. T., Close, K. L., et al.** Clinical targets for continuous glucose monitoring data interpretation: recommendations from the international consensus on time in range. *Diabetes Care 42,* 8 (2019), 1593–1603.

[12] **Bay, S. D., and Pazzani, M. J.** Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery 5* (2001), 213–246.

[13] **Bayardo, R. J., Agrawal, R., and Gunopulos, D.** Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery 4* (2000), 217–240.

[14] **Becker, M., Lemmerich, F., Singer, P., Strohmaier, M., and Hotho, A.** MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data. *Data Mining and Knowledge Discovery 31*, 5 (2017), 1359–1390.

[15] **Belfodil, A., Belfodil, A., and Kaytoue, M.** Anytime subgroup discovery in numerical domains with guarantees. In *Proc. ECML PKDD* (2018), pp. 500–516.

[16] **Belfodil, A., Cazalens, S., Lamarre, P., and Plantevit, M.** Flash points: Discovering exceptional pairwise behaviors in vote or rating data. In *Proc. ECML PKDD* (2017), pp. 442–458.

[17] **Belfodil, A., Duivesteijn, W., Plantevit, M., Cazalens, S., and Lamarre, P.** DEvIANT: Discovering significant exceptional (dis-)agreement within groups. In *Proc. ECML PKDD* (2020), pp. 3–20.

[18] **Benavides-Varela, S., Laurillard, D., Piperno, G., Fava Minor, D., Lucangeli, D., and Butterworth, B.** Digital games for learning basic arithmetic at home. In *Game-Based Learning in Education and Health - Part A*, F. H. Santos, Ed. Elsevier, 2023, pp. 35–61.

[19] **Bendimerad, A., Lijffijt, J., Plantevit, M., Robardet, C., and De Bie, T.** Contrastive antichains in hierarchies. In *Proc. KDD* (2019), pp. 294–304.

[20] **Bendimerad, A., Plantevit, M., and Robardet, C.** Mining exceptional closed patterns in attributed graphs. *Knowledge and Information Systems 56* (2018), 1–25.

[21] **Bendimerad, A. A., Plantevit, M., and Robardet, C.** Unsupervised exceptional attributed sub-graph mining in urban data. In *Proc. ICDM* (2016), pp. 21–30.

[22] **Bethlehem, J.** *Applied survey methods: A statistical perspective.* John Wiley & Sons, 2009.

[23] **Bishop, C. M.** *Pattern recognition and machine learning.* Springer, 2006.

[24] **Bosc, G., Boulicaut, J.-F., Raïssi, C., and Kaytoue, M.** Anytime discovery of a diverse set of patterns with Monte Carlo Tree Search. *Data Mining and Knowledge Discovery 32*, 3 (2018), 604–650.

[25] **Bryman, A.** *Social Research Methods.* Oxford University Press, 2016.

[26] **Bueno, M. L., Hommersom, A., and Lucas, P. J.** Temporal Exceptional Model Mining using Dynamic Bayesian Networks. In *Proc. 5th ECML PKDD Workshop AALTD* (2020), pp. 97–112.

[27] **Bueno, M. L., Hommersom, A., Lucas, P. J., and Janzing, J.** A probabilistic framework for predicting disease dynamics: A case study of psychotic depression. *Journal of Biomedical Informatics 95* (2019), 103232.

[28] **Burnham, K. P., and Anderson, D. R.** Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research 33*, 2 (2004), 261–304.

[29] **Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., and García, S.** Subgroup discovery applied to the e-commerce website OrOliveSur.com. In *ICEIS (2)* (2012), pp. 239–244.

[30] **Centraal Bureau van de Statistiek (Statistics Netherlands)**. Jaarrapport 2016 Landelijke Jeugdmonitor (Annual report 2016 National Youth Monitor). Tech. rep., CBS, 2016. Available through `https://www.cbs.nl/nl-nl/publicatie/2016/48/jaarrapport-2016-landelijke-jeugdmonitor`.

[31] **Chen, P. P.-S.** The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems 1*, 1 (1976), 9–36.

[32] **Clark, P., and Niblett, T.** The CN2 induction algorithm. *Machine Learning 3*, 4 (1989), 261–283.

[33] **Codd, E. F.** A relational model of data for large shared data banks. *Communications of the ACM 13*, 6 (1970), 377–387.

[34] **Cohen, W. W.** Fast effective rule induction. In *Proc. Machine Learning*. Elsevier, 1995, pp. 115–123.

[35] **Cole, E. R.** Intersectionality and research in psychology. *American psychologist 64*, 3 (2009), 170.

[36] **Crenshaw, K.** Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum 1989*, 1 (1989), 139–167.

[37] **Dagum, P., Galper, A., and Horvitz, E.** Dynamic network models for forecasting. In *Proc. UAI* (1992), pp. 41–48.

[38] **Danne, T., Nimri, R., Battelino, T., Bergenstal, R. M., Close, K. L., DeVries, J. H., Garg, S., Heinemann, L., Hirsch, I., Amiel, S. A., et al.** International consensus on use of continuous glucose monitoring. *Diabetes Care 40*, 12 (2017), 1631–1640.

[39] **de Looze, M. E., Raaijmakers, Q., ter Bogt, T. F., Bendtsen, P., Farhat, T., Ferreira, M., Godeau, E., Kuntsche, E., Molcho, M., Pförtner, T.-K., et al.** Decreases in adolescent weekly alcohol use in Europe and North America: Evidence from 28 countries from 2002 to 2010. *The European Journal of Public Health 25*, 2 (2015), 69–72.

[40] **de Looze, M. E., van Dorsselaer, S. A., Monshouwer, K., and Vollebergh, W. A.** Trends in adolescent alcohol use in the Netherlands, 1992–2015: Differences across sociodemographic groups and links with strict parental rule-setting. *International Journal of Drug Policy 50* (2017), 90–101.

[41] **de Looze, M. E., Vermeulen-Smit, E., ter Bogt, T. F., van Dorsselaer, S. A., Verdurmen, J., Schulten, I., Engels, R. C., and Vollebergh, W. A.** Trends in alcohol-specific parenting practices and adolescent alcohol use between 2007 and 2011 in the Netherlands. *International Journal of Drug Policy 25*, 1 (2014), 133–141.

[42] **de Sá, C. R., Duivesteijn, W., Azevedo, P., Jorge, A. M., Soares, C., and Knobbe, A.** Discovering a taste for the unusual: Exceptional models for preference mining. *Machine Learning 107* (2018), 1775–1807.

[43] **de Vries, C. P., Schouten, R. M., van der Kuur, J., Gottardi, L., and Akamatsu, H.** Microcalorimeter pulse analysis by means of principle component decomposition. In *Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray* (2016), vol. 99055V, pp. 1699–1708.

[44] **de Vries, H., Dijkstra, M., and Kuhlman, P.** Self-efficacy: The third factor besides attitude and subjective norm as a predictor of behavioural intentions. *Health Education Research 3*, 3 (1988), 273–282.

[45] **Dekking, F. M., Kraaikamp, C., Lopuhaä, H., and Meester, L.** *A modern introduction to probability and statistics: Understanding why and how.* Springer Science & Business Media, 2005.

[46] **den Braber, N., Vollenborek-Hutten, M. M., Westerik, K. M., Bakker, S. J., Navis, G., van Beijnum, B.-J. F., and Laverman, G. D.** Glucose regulation beyond HbA1c in type 2 diabetes treated with insulin: Real-world evidence from the DIALECT-2 cohort. *Diabetes Care 44*, 10 (2021), 2238–2244.

[47] **Deng, J., Kang, B., Lijffijt, J., and Bie, T. D.** Explainable subgraphs with surprising densities: A subgroup discovery approach. In *Proc. SDM* (2020), pp. 586–594.

[48] **Dong, G., and Li, J.** Efficient mining of emerging patterns: Discovering trends and differences. In *Proc. KDD* (1999), pp. 43–52.

[49] **Dowd, J. J., and Bengtson, V. L.** Aging in minority populations: An examination of the double jeopardy hypothesis. *Journal of Gerontology 33*, 3 (1978), 427–436.

[50] **Du, X., Pei, Y., Duivesteijn, W., and Pechenizkiy, M.** Exceptional spatio-temporal behavior mining through Bayesian non-parametric modeling. *Data Mining and Knowledge Discovery 34* (2020), 1267–1290.

[51] **Du, X., Pei, Y., Duivesteijn, W., and Pechenizkiy, M.** Fairness in network representation by latent structural heterogeneity in observational data. *Proc. AAAI 34*, 04 (2020), 3809–3816.

[52] **Duivesteijn, W., Farzami, T., Putman, T., Peer, E., Weerts, H. J., Adegeest, J. N., Foks, G., and Pechenizkiy, M.** Have it both ways — from A/B testing to A&B testing with Exceptional Model Mining. In *Proc. ECML PKDD* (2017), pp. 114–126.

[53] **Duivesteijn, W., Feelders, A., and Knobbe, A.** Different slopes for different folks: Mining for exceptional regression models with Cook's distance. In *Proc. KDD* (2012), pp. 868–876.

[54] **Duivesteijn, W., Feelders, A. J., and Knobbe, A.** Exceptional Model Mining. *Data Mining and Knowledge Discovery 30*, 1 (2016), 47–98.

[55] **Duivesteijn, W., and Knobbe, A.** Exploiting false discoveries: Statistical validation of patterns and quality measures in subgroup discovery. In *Proc. ICDM* (2011), pp. 151–160.

[56] **Duivesteijn, W., Knobbe, A., Feelders, A., and van Leeuwen, M.** Subgroup discovery meets Bayesian networks: An Exceptional Model Mining approach. In *Proc. ICDM* (2010), pp. 158–167.

[57] **Fani Sani, M., van der Aalst, W., Bolt, A., and García-Algarra, J.** Subgroup discovery in process mining. In *Proc. BIS* (2017), vol. 288, p. 237 – 252.

[58] **Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P.** Knowledge discovery and data mining: Towards a unifying framework. In *Proc. KDD* (1996), pp. 82–88.

[59] **Feigenson, L., Dehaene, S., and Spelke, E.** Core systems of number. *Trends in cognitive sciences 8*, 7 (2004), 307–314.

[60] **Flach, P., and Lachiche, N.** 1BC: A first-order Bayesian classifier. In *Proc. ICILP* (1999), pp. 92–103.

[61] **Galbrun, E., and Kimmig, A.** Towards finding relational redescriptions. In *Proc. DS* (2012), pp. 52–66.

[62] **Galbrun, E., and Kimmig, A.** Finding relational redescriptions. *Machine learning 96* (2014), 225–248.

[63] **Galbrun, E., and Miettinen, P.** From black and white to full color: Extending redescription mining outside the Boolean world. *Statistical Analysis and Data Mining: The ASA Data Science Journal 5*, 4 (2012), 284–303.

[64] **Gallo, A., Miettinen, P., and Mannila, H.** Finding subgroups having several descriptions: Algorithms for redescription mining. In *Proc. SDM* (2008), pp. 334–345.

[65] **Gamberger, D., and Lavrac, N.** Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research 17* (2002), 501–527.

[66] **Gamberger, D., Lučanin, D., and Šmuc, T.** Analysis of world bank indicators for countries with banking crises by subgroup discovery induction. In *Proc. MIPRO)* (2013), pp. 1138–1142.

[67] **Gant, C. M., Binnenmars, S. H., van den Berg, E., Bakker, S. J., Navis, G., and Laverman, G. D.** Integrated assessment of pharmacological and nutritional cardiovascular risk management: Blood pressure control in the DIAbetes and LifEstyle Cohort Twente (DIALECT). *Nutrients 9*, 7 (2017), 709.

[68] **Geels, L. M., Bartels, M., van Beijsterveldt, T. C., Willemsen, G., van der Aa, N., Boomsma, D. I., and Vink, J. M.** Trends in adolescent alcohol use: Effects of age, sex and cohort on prevalence and heritability. *Addiction 107*, 3 (2012), 518–527.

[69] **Ghavami, N., Katsiaficas, D., and Rogers, L. O.** Toward an intersectional approach in developmental science: The role of race, gender, sexual orientation, and immigrant status. *Advances in child development and behavior 50* (2016), 31–73.

[70] **Gneiting, T., and Raftery, A. E.** Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*, 477 (2007), 359–378.

[71] **Gray, S. A., and Reeve, R. A.** Preschoolers' dot enumeration abilities are markers of their arithmetic competence. *PLoS One 9*, 4 (2014), e94428.

[72] **Grosskreutz, H., Boley, M., and Krause-Traudes, M.** Subgroup discovery for election analysis: A case study in descriptive data mining. In *Proc. DS* (2010), pp. 57–71.

[73] **Han, J., Cheng, H., Xin, D., and Yan, X.** Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery 15*, 1 (2007), 55–86.

[74] **Hand, D. J.** Pattern detection and discovery. In *Proc. Pattern Detection and Discovery: ESF Exploratory Workshop London* (2002), pp. 1–12.

[75] **Hedeker, D., Mermelstein, R. J., and Demirtas, H.** An application of a mixed-effects location scale model for analysis of Ecological Momentary Assessment (EMA) data. *Biometrics 64*, 2 (2008), 627–634.

[76] **Helal, S.** Subgroup discovery algorithms: A survey and empirical evaluation. *Journal of Computer Science and Technology 31* (2016), 561–576.

[77] **Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., and Murray, D. J.** Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics 7* (2019), 227–245.

[78] **Hellstrand, H., Holopainen, S., Korhonen, J., Räsänen, P., Hakkarainen, A., Laakso, M.-J., Laine, A., and Aunio, P.** Arithmetic fluency and number processing skills in identifying students with mathematical learning disabilities. *Available at SSRN 4641604* (2023).

[79] **Hermens, H., op den Akker, H., Tabak, M., Wijsman, J., and Vollenbroek, M.** Personalized coaching systems to support healthy behavior in people with chronic conditions. *Journal of Electromyography and Kinesiology 24*, 6 (2014), 815–826.

[80] **Herrera, F., Carmona, C. J., González, P., and Del Jesus, M. J.** An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems 29*, 3 (2011), 495–525.

[81] **Hinder, F., Artelt, A., Vaquet, V., and Hammer, B.** Contrasting explanation of concept Drift. In *Proc. ESANN* (2022).

[82] **Hobolt, S. B., Leeper, T. J., and Tilley, J.** Divided by the vote: Affective polarization in the wake of the Brexit referendum. *British Journal of Political Science 51*, 4 (2021), 1476–1493.

[83] **Holloway, I. D., and Ansari, D.** Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology 103*, 1 (2009), 17–29.

[84] **Hox, J., Moerbeek, M., and van de Schoot, R.** *Multilevel analysis: Techniques and applications.* Routledge, 2017.

[85] **Hudson, D., Wiltshire, T. J., and Atzmueller, M.** Local exceptionality detection in time series using subgroup discovery: An approach exemplified on team interaction data. In *Proc. DS* (2021), pp. 435–445.

[86] **Hudson, D., Wiltshire, T. J., and Atzmueller, M.** Visualization methods for exploratory subgroup discovery on time series data. In *Proc. IWINAC* (2022), pp. 34–44.

[87] **Hurvich, C. M., and Tsai, C.-L.** Model selection for extended quasi-likelihood models in small samples. *Biometrics 51*, 3 (1995), 1077–1084.

[88] **Iferroudjene, M., Lonjarret, C., Robardet, C., Plantevit, M., and Atzmueller, M.** Methods for explaining top-N recommendations through subgroup discovery. *Data Mining and Knowledge Discovery 37*, 2 (2023), 833–872.

[89] **IJsselhof, R. J., Duchateau, S. D., Schouten, R. M., Freund, M. W., Heuser, J., Fejzic, Z., Haas, F., Schoof, P. H., and Slieker, M. G.** Follow-up after biventricular repair of the hypoplastic left heart complex. *European Journal of Cardio-Thoracic Surgery 57*, 4 (2019), 644–651.

[90] **IJsselhof, R. J., Duchateau, S. D., Schouten, R. M., Slieker, M. G., Hazekamp, M. G., and Schoof, P. H.** Long-term follow-up of pericardium for the ventricular component in atrioventricular septal defect repair. *World Journal for Pediatric and Congenital Heart Surgery 11*, 6 (2020), 742–747.

[91] **Jaroszewicz, S.** Using interesting sequences to interactively build Hidden Markov Models. *Data Mining and Knowledge Discovery 21*, 1 (2010), 186–220.

[92] **Jekel, C. F., and Venter, G.** pwlf: A python library for fitting 1D continuous piecewise linear functions, 2019. `https://github.com/cjekel/piecewise_linear_fit_py`.

[93] **Jin, N., Flach, P., Wilcox, T., Sellman, R., Thumim, J., and Knobbe, A.** Subgroup discovery in smart electricity meter data. *IEEE Transactions on Industrial Informatics 10*, 2 (2014), 1327–1336.

[94] **Johnston, L. D., Miech, R. A., O'Malley, P. M., Bachman, J. G., Schulenberg, J. E., and Patrick, M. E.** *Monitoring the Future national survey results on drug use 1975-2020: Overview, key findings on adolescent drug use.* Ann Arbor: Institute for Social Research, University of Michigan, 2021.

[95] **Kalofolias, J., and Vreeken, J.** Naming the most anomalous cluster in Hilbert space for structures with attribute information. *Proc. AAAI 36*, 4 (2022), 4057–4064.

[96] **Kappen, I. F., Bittermann, G. K., Schouten, R. M., Bittermann, D., Etty, E., Koole, R., Kon, M., Mink van der Molen, A., and Breugem, C. C.** Long-term mid-facial growth of patients with a unilateral complete cleft of lip, alveolus and palate treated by two-stage palatoplasty: Cephalometric analysis. *Clinical Oral Investigations 21* (2017), 1801–1810.

[97] **Kass, R. E., and Raftery, A. E.** Bayes Factors. *Journal of the American Statistical Association 90*, 430 (1995), 773–795.

[98] **Kaytoue, M., Plantevit, M., Zimmermann, A., Bendimerad, A., and Robardet, C.** Exceptional contextual subgraph mining. *Machine Learning 106* (2017), 1171–1211.

[99] **Kern, M. R., Duinhof, E. L., Walsh, S. D., Cosma, A., Moreno-Maldonado, C., Molcho, M., Currie, C., and Stevens, G. W.** Intersectionality and adolescent mental well-being: A cross-nationally comparative analysis of the interplay between immigration background, socioeconomic status and gender. *Journal of Adolescent Health 66*, 6 (2020), S12–S20.

[100] **King, D. K.** Multiple jeopardy, multiple consciousness: The context of a black feminist ideology. *Signs: Journal of women in culture and society 14*, 1 (1988), 42–72.

[101] **Kiseleva, J., Lam, H. T., Pechenizkiy, M., and Calders, T.** Predicting current user intent with contextual Markov models. In *Proc. ICDM Workshops* (2013), pp. 391–398.

[102] **Klösgen, W.** Explora: A multipattern and multistrategy discovery assistant. In *Proc. PAKDD* (1996), pp. 249–271.

[103] **Klösgen, W., and May, M.** Census data mining — An application. In *Proc. PKDD* (2002), pp. 65–79.

[104] **Knobbe, A., Crémilleux, B., Fürnkranz, J., and Scholz, M.** From local patterns to global models: the LeGo approach to data mining. In *From local patterns to global models: Proceedings of the ECML/PKDD-08 workshop (LeGo-08)* (2008), University and State Library Darmstadt Darmstadt, Germany, pp. 1–16.

[105] **Knobbe, A. J., and Ho, E. K. Y.** Pattern teams. In *Proc. ECML PKDD* (2006), pp. 577–584.

[106] **Konijn, R. M., Duivesteijn, W., Meeng, M., and Knobbe, A.** Cost-based quality measures in subgroup discovery. *Journal of Intelligent Information Systems 45* (2015), 337–355.

[107] **Konijn, R. M., and Kowalczyk, W.** Hunting for fraudsters in random forests. In *Proc. HAIS* (2012), pp. 174–185.

[108] **Kovatchev, B. P., Cox, D. J., Kumar, A., Gonder-Frederick, L., and Clarke, W. L.** Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data. *Diabetes Technology & Therapeutics 5,* 5 (2003), 817–828.

[109] **Krak, T. E., and Feelders, A.** Exceptional Model Mining with tree-constrained gradient ascent. In *Proc. SDM* (2015), pp. 487–495.

[110] **Kralj Novak, P., Lavrač, N., and Webb, G. I.** Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research 10,* 2 (2009).

[111] **Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A.** Outlier detection in arbitrarily oriented subspaces. In *Proc. ICDM* (2012), pp. 379–388.

[112] **Krogel, M.-A., and Wrobel, S.** Transformation-based learning using multirelational aggregation. In *Proc. ILP* (2001), pp. 142–155.

[113] **Lakhal, L., and Stumme, G.** Efficient mining of association rules based on formal concept analysis. In *Formal concept analysis: Foundations and Applications.* Springer, 2005, pp. 180–195.

[114] **Lambach, D., and Gamberger, D.** Temporal analysis of political instability through descriptive subgroup discovery. *Conflict Management and Peace Science 25,* 1 (2008), 19–32.

[115] **Landerl, K., Bevan, A., and Butterworth, B.** Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition 93,* 2 (2004), 99–125.

[116] **Lavrač, N., Džeroski, S., and Grobelnik, M.** Learning nonrecursive definitions of relations with LINUS. In *Proc. EWSL* (1991), pp. 265–281.

[117] **Lavrač, N., Kavšek, B., Flach, P., and Todorovski, L.** Subgroup discovery with CN2-SD. *Journal of Machine Learning Research 5,* 2 (2004), 153–188.

[118] **Lavrac, N., Novak, P. K., Mozetic, I., Podpecan, V., Motaln, H., Petek, M., and Gruden, K.** Semantic subgroup discovery: Using ontologies in microarray data analysis. In *Proc. IEEE Engineering in Medicine and Biology Society* (2009), pp. 5613–5616.

[119] **Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I., and Novak, P. K.** Using ontologies in semantic data mining with SEGS and g-SEGS. In *Proc. DS* (2011), pp. 165–178.

[120] **Leibovich-Raveh, T., Lewis, D. J., Kadhim, S. A.-R., and Ansari, D.** A new method for calculating individual subitizing ranges. *Journal of Numerical Cognition 4,* 2 (2018), 429–447.

[121] **Leman, D., Feelders, A., and Knobbe, A.** Exceptional Model Mining. In *Proc. ECML PKDD* (2008), pp. 1–16.

[122] **Lemmerich, F., Atzmueller, M., and Puppe, F.** Fast exhaustive subgroup discovery with numerical target concepts. *Data Mining and Knowledge Discovery 30,* 3 (2016), 711–762.

[123] **Lemmerich, F., Becker, M., and Atzmueller, M.** Generic pattern trees for exhaustive Exceptional Model Mining. In *Proc. ECML PKDD* (2012), pp. 277–292.

[124] **Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., and Strohmaier, M.** Mining subgroups with exceptional transition behavior. In *Proc. KDD* (2016), pp. 965–974.

[125] **Leyland, A. H., and Groenewegen, P. P.** *Multilevel modelling for public health and health services research: Health in context.* Springer Nature, 2020.

[126] **Li, X., and Han, J.** Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proc. VLDB* (2007), pp. 447–458.

[127] **Lijffijt, J., Kang, B., Duivesteijn, W., Puolamaki, K., Oikarinen, E., and De Bie, T.** Subjectively interesting subgroup discovery on real-valued targets. In *Proc. ICDE* (2018), pp. 1352–1355.

[128] **Little, R. J. A., and Rubin, D. B.** *Statistical analysis with missing data,* vol. 793. John Wiley & Sons, 2019.

[129] **Lonjarret, C., Robardet, C., Plantevit, M., Auburtin, R., and Atzmueller, M.** Why should I trust this item? Explaining the recommendations of any model. In *Proc. DSAA* (2020), pp. 526–535.

[130] **Major, C. S., Paul, J. M., and Reeve, R. A.** TEMA and dot enumeration profiles predict mental addition problem solving speed longitudinally. *Frontiers in Psychology 8* (2017), 313803.

[131] **Mathonat, R., Nurbakova, D., Boulicaut, J.-F., and Kaytoue, M.** Anytime mining of sequential discriminative patterns in labeled sequences. *Knowledge and Information Systems 63,* 2 (2021), 439–476.

[132] **Mathonat, R., Nurbakova, D., Boulicaut, J.-F., and Kaytoue, M.** Anytime subgroup discovery in high dimensional numerical data. In *Proc. DSAA* (2021), pp. 1–10.

[133] **McCambridge, J., McAlaney, J., and Rowe, R.** Adult consequences of late adolescent alcohol consumption: A systematic review of cohort studies. *PLoS Medicine 8,* 2 (2011), e1000413.

[134] **McGuire, H., Longson, D., Adler, A., Farmer, A., and Lewin, I.** Management of type 2 diabetes in adults: Summary of updated NICE guidance. *BMJ 353* (2016).

[135] **Meeng, M., and Knobbe, A. J.** For real: A thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery 35,* 1 (2021), 158–212.

[136] **Meier, J., Dietz, A., Boehm, A., and Neumuth, T.** Predicting treatment process steps from events. *Journal of Biomedical Informatics 53* (2015), 308–319.

[137] **Mokinaro, S., Vincente, J., Benedetti, E., Cerrai, S., Colasante, E., Arpa, S., Chomynova, P., Kraus, L., Monshouwer, K., Spika, S., et al.** *ESPAD Report 2019: Results from European School survey Project on Alcohol and other Drugs.* Technological University Dublin, 2020.

[138] **Mollenhauer, D., and Atzmueller, M.** Sequential exceptional pattern discovery using pattern-growth: An extensible framework for interpretable machine learning on sequential data. In *Proc. XI-ML* (2020).

[139] **Morik, K., Boulicaut, J.-F., and Siebes, A.** *Local pattern detection: International Seminar Dagstuhl Castle, Germany, April 12-16, 2004, Revised Selected Papers*, vol. 3539. Springer, 2005.

[140] **Mozetič, I., Lavrač, N., Podpečan, V., Novak, P. K., Motaln, H., Petek, M., Gruden, K., Toivonen, H., and Kulovesi, K.** Semantic subgroup discovery and cross-context linking for microarray data analysis. *Bisociative Knowledge Discovery: An Introduction to Concept, Algorithms, Tools, and Applications* (2012), 379–389.

[141] **Muggeo, V. M.** Estimating regression models with unknown break-points. *Statistics in Medicine 22*, 19 (2003), 3055–3071.

[142] **Muggleton, S.** Inverse entailment and Progol. *New generation computing 13* (1995), 245–286.

[143] **Mulders, P. J., van den Heuvel, E. R., Reidsma, P., and Duivesteijn, W.** Introducing exceptional growth mining – analyzing the impact of soil characteristics on on-farm crop growth and yield variability. *PLoS ONE 19*, 1 (2024), e0296684.

[144] **Muschalik, M., Fumagalli, F., Hammer, B., and Hüllermeier, E.** Agnostic explanation of model change based on feature importance. *Künstliche Intelligenz 36*, 3 (2022), 211–224.

[145] **op den Akker, H., Cabrita, M., op den Akker, R., Jones, V. M., and Hermens, H. J.** Tailored motivational message generation: A model and practical framework for real-time physical activity coaching. *Journal of Biomedical Informatics 55* (2015), 104–115.

[146] **op den Akker, H., Jones, V. M., and Hermens, H. J.** Tailoring real-time physical activity coaching systems: A literature survey and model. *User modeling and user-adapted interaction 24* (2014), 351–392.

[147] **Pastor, E., Baralis, E., and de Alfaro, L.** A hierarchical approach to anomalous subgroup discovery. In *Proc. ICDE* (2023), pp. 2647–2659.

[148] **Peharz, R., Kapeller, G., Mowlaee, P., and Pernkopf, F.** Modeling speech with sum-product networks: Application to bandwidth extension. In *Proc. ICASSP* (2014), pp. 3699–3703.

[149] **Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C.** Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering 16*, 11 (2004), 1424–1440.

[150] **Pieters, B. F., Knobbe, A., and Dzeroski, S.** Subgroup discovery in ranked data, with an application to gene set enrichment. In *Proc. Preference Learning Workshop at ECML PKDD* (2010), pp. 1–18.

[151] **Pirolli, P. L., and Pitkow, J. E.** Distributions of surfers' paths through the World Wide Web: Empirical characterizations. *World Wide Web 2*, 1-2 (1999), 29–45.

[152] **Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M.** Selecting the number of states in Hidden Markov Models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics 22*, 3 (2017), 270–293.

[153] **Proença, H. M., Grünwald, P., Bäck, T., and van Leeuwen, M.** Discovering outstanding subgroup lists for numeric targets using MDL. In *Proc. ECML PKDD* (2020).

[154] **Ramakrishnan, N., Kumar, D., Mishra, B., Potts, M., and Helm, R. F.** Turning CARTwheels: An alternating algorithm for mining redescriptions. In *Proc. SDM* (2004), pp. 266–275.

[155] **Räsänen, P., Aunio, P., Laine, A., Hakkarainen, A., Väisänen, E., Finell, J., Rajala, T., Laakso, M.-J., and Korhonen, J.** Effects of gender on basic numerical and arithmetic skills: Pilot data from third to ninth grade for a large-scale online dyscalculia screener. *Frontiers in Education: Educational Psychology 6* (2021), 683672.

[156] **Rebelo de Sá, C., Duivesteijn, W., Soares, C., and Knobbe, A.** Exceptional preferences mining. In *Proc. DS* (2016), pp. 3–18.

[157] **Reeve, R., Reynolds, F., Humberstone, J., and Butterworth, B.** Stability and change in markers of core numerical competencies. *Journal of Experimental Psychology: General 141*, 4 (2012), 649.

[158] **Remil, Y., Bendimerad, A., Chambard, M., Mathonat, R., Plantevit, M., and Kaytoue, M.** Mining java memory errors using subjective interesting subgroups with hierarchical targets. In *Proc. ICDM* (2023), pp. 1221–1230.

[159] **Ribeiro, J., Fontes, T., Soares, C., and Borges, J.** Multidimensional subgroup discovery on event logs. *Expert Systems with Applications 246* (2024), 123205.

[160] **Richter, M., Kuntsche, E., de Looze, M., and Pförtner, T.-K.** Trends in socioeconomic inequalities in adolescent alcohol use in Germany between 1994 and 2006. *International journal of public health 58*, 5 (2013), 777–784.

[161] **Rombouts, M., van Dorsselaer, S. A., Scheffers-van Schayck, T., Tuithof, M., et al.** *Jeugd en riskant gedrag 2019. Kerngegevens uit het Peilstationsonderzoek Scholieren.* Trimbos-Instituut, Utrecht, 2020.

[162] **Rubin, D. B.** Inference and missing data. *Biometrika 63*, 3 (1976), 581–592.

[163] **Rudaś, K., and Jaroszewicz, S.** Linear regression for uplift modeling. *Data Mining and Knowledge Discovery 32*, 5 (2018), 1275–1305.

[164] **Sadagopan, N., and Li, J.** Characterizing typical and atypical user sessions in clickstreams. In *Proc. WWW* (2008), pp. 885–894.

[165] **Sarukkai, R. R.** Link prediction and path analysis using Markov chains. *Computer Networks 33*, 1-6 (2000), 377–386.

[166] **Schek, H.-J., and Scholl, M. H.** The relational model with relation-valued attributes. *Information Systems 11*, 2 (1986), 137–147.

[167] **Schmitt, H., Scholz, E., Leim, I., and Moschner, M.** The Mannheim Eurobarometer trend file 1970-2002. GESIS Data Archive, Cologne. ZA3521 Data file Version 2.0.1, 2008.

[168] **Schoof, J. T., and Pryor, S.** On the proper order of Markov chain model for daily precipitation occurrence in the contiguous United States. *Journal of Applied Meteorology and Climatology 47*, 9 (2008), 2477–2486.

[169] **Schouten, R. M.** On the role of prognostic factors and effect modifiers in structural causal models. *Accepted for poster presentation at Causal Representation Learning Workshop NeurIPS* (2024).

[170] **Schouten, R. M., Bueno, M. L., Duivesteijn, W., and Pechenizkiy, M.** Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery 36* (2022), 379–413.

[171] **Schouten, R. M., Duivesteijn, W., and Pechenizkiy, M.** Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS). In *Proc. SDM* (2022), pp. 585–593.

[172] **Schouten, R. M., Duivesteijn, W., and Pechenizkiy, M.** Exceptional model mining for repeated cross-sectional data (EMM-RCS) — supplementary material. Tech. rep., Available at Figshare: `https://doi.org/10.6084/m9.figshare.18688220/`, 2022.

[173] **Schouten, R. M., Duivesteijn, W., Räsänen, P., Paul, J. M., and Pechenizkiy, M.** Exceptional Subitizing Patterns — Supplementary Material. Tech. rep., available at Figshare, `https://doi.org/10.6084/m9.figshare.26008879`, 2024.

[174] **Schouten, R. M., Duivesteijn, W., Räsänen, P., Paul, J. M., and Pechenizkiy, M.** Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression. In *Proc. ECML PKDD* (2024), p. 66–82.

[175] **Schouten, R. M., Engelen, B. L., Duivesteijn, W., and Pechenizkiy, M.** Towards a unified framework for Exceptional Model Mining for hierarchical data. *To be submitted to IJCAI 2025* (2024).

[176] **Schouten, R. M., Lugtig, P., and Vink, G.** Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation 88*, 15 (2018), 2909–2930.

[177] **Schouten, R. M., Stevens, G. W., van Dorsselaer, S. A., Duinhof, E. L., Monshouwer, K., Pechenizkiy, M., and Duivesteijn, W.** Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining: Discovering exceptional trends in adolescent alcohol use in the Netherlands. *Accepted for presentation at BNAIC/BeNeLearn* (2024).

[178] **Schouten, R. M., Taşcău, V., Ziegler, G. G., Casano, D., Ardizzone, M., and Erotokritou, M. A.** Dropping incomplete records is (not so) straightforward. In *Proc. IDA* (2023), pp. 379–391.

[179] **Schouten, R. M., and Vink, G.** The dance of the mechanisms: How observed information influences the validity of missingness assumptions. *Sociological Methods & Research 50* (2021), 1243–1258.

[180] **Schwarz, G.** Estimating the dimension of a model. *The Annals of Statistics 6*, 2 (1978), 461–464.

[181] **Shalev, R. S., Auerbach, J., Manor, O., and Gross-Tsur, V.** Developmental dyscalculia: prevalence and prognosis. *European child & adolescent psychiatry 9* (2000), S58–S64.

[182] **Siebes, A.** Data surveying: Foundations of an inductive query language. In *Proc. KDD* (1995), pp. 269–274.

[183] **Simpson, E. H.** The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological) 13*, 2 (1951), 238–241.

[184] **Singer, P., Helic, D., Taraghi, B., and Strohmaier, M.** Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *PloS one 9*, 7 (2014), e102070.

[185] **Škrlj, B., Kralj, J., Vavpetič, A., and Lavrač, N.** Community-based semantic subgroup discovery. In *Proc. New Frontiers in Mining Complex Patterns* (2018), pp. 182–196.

[186] **Sleet, D. A., Ballesteros, M. F., and Borse, N. N.** A review of unintentional injuries in adolescents. *Annual Review of Public Health 31* (2010), 195–212.

[187] **Snijders, T., and Bosker, R.** *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publishers, 2012.

[188] **Song, H.** *Model-based subgroup discovery.* PhD thesis, University of Bristol, 2017.

[189] **Song, H., Flach, P., and Kalogridis, G.** Dataset shift detection with model-based subgroup discovery. In *Proc. LMCE* (2015).

[190] **Song, H., Kull, M., Flach, P., and Kalogridis, G.** Subgroup discovery with proper scoring rules. In *Proc. ECML PKDD* (2016), pp. 492–510.

[191] **Stamm, F. I., Becker, M., Strohmaier, M., and Lemmerich, F.** Redescription model mining. In *Proc. KDD* (2021), pp. 1521–1529.

[192] **Stevens, G. W. J. M., Van Dorsselaer, S., Boer, M., De Roos, S., Duinhof, E., Ter Bogt, T., Van Den Eijnden, R., Kuyper, L., Visser, D., Vollebergh, W. A. M., et al.** *HBSC 2017. Gezondheid en welzijn van jongeren in Nederland.* Utrecht University, 2018.

[193] **Sugiura, N.** Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods 7*, 1 (1978), 13–26.

[194] **Suzuki, E.** Undirected exception rule discovery as local pattern detection. In *Proc. Local Pattern Detection: International Seminar* (2005), pp. 207–216.

[195] **Tong, H.** Determination of the order of a Markov chain by Akaike's information criterion. *Journal of Applied Probability 12*, 3 (1975), 488–497.

[196] **van Buuren, S.** *Flexible imputation of missing data*, 2nd ed. Chapman & Hall/CRC: Boca Raton, 2018.

[197] **van den Berg, N. T., Broekgaarden, B. O., Mahieu Dionysia, P., Martens, J. G., Niederle, J., Schouten, R. M., and Duivesteijn, W.** Generating MNAR missingness in image data, with additional evaluation of MisGAN. *Accepted for presentation at BNAIC* (2024).

[198] **van der Haar, J. F., Nagelkerken, S. C., Smit, I. G., van Straaten, K., Tack, J. A., Schouten, R. M., and Duivesteijn, W.** Efficient Subgroup Discovery through Auto-Encoding. In *Proc. IDA* (2022), pp. 327–340.

[199] **van Giessen, A., Boer, J., van Gestel, I., Douma, E., du Pon, E., Blokstra, A., and Koopman, N.** *Voortgangsrapportage Nationaal Preventieakkoord 2019.* RIVM, 2020.

[200] **van Leeuwen, M., and Knobbe, A.** Non-redundant subgroup discovery in large and complex data. In *Proc. ECML PKDD* (2011), pp. 459–474.

[201] **van Leeuwen, M., and Knobbe, A.** Diverse subgroup set discovery. *Data Mining and Knowledge Discovery 25*, 2 (2012), 208–242.

[202] **Vavpetič, A., and Lavrač, N.** Semantic subgroup discovery systems and workflows in the SDM-toolkit. *The Computer Journal 56*, 3 (2013), 304–320.

[203] **Vavpetič, A., Podpečan, V., and Lavrač, N.** Semantic subgroup explanations. *Journal of Intelligent Information Systems 42* (2014), 233–254.

[204] **Verhaegh, R. F. A., Kiezebrink, J. J. E., Nusteling, F., Rio, A. W. A., Bendicsek, M. B., Duivesteijn, W., and Schouten, R. M.** A clustering-inspired quality measure for exceptional preferences mining – design choices and consequences. In *Proc. DS* (2022), pp. 429–444.

[205] **Webb, G. I.** Discovering significant patterns. *Machine learning 68* (2007), 1–33.

[206] **Webb, G. I., Butler, S., and Newlands, D.** On detecting differences between groups. In *Proc. KDD* (2003), pp. 256–265.

[207] **Wilks, D. S.** Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology 93*, 3 (1999), 153–169.

[208] **World Health Organization**. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: Abbreviated report of a WHO consultation. No. WHO/NMH/CH-P/CPM/11.1, WHO, 2011.

[209] **Wrobel, S.** An algorithm for multi-relational discovery of subgroups. In *Proc. PKDD* (1997), pp. 78–87.

[210] **Wrobel, S.** Inductive logic programming for knowledge discovery in databases. In *Relational data mining*. Springer, 2001, pp. 74–101.

[211] **Železnỳ, F., and Lavrač, N.** Propositionalization-based relational subgroup discovery with RSD. *Machine Learning 62* (2006), 33–63.

[212] **Zimek, A., and Vreeken, J.** The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning (ML) 98* (2015), 121–155.

[213] **Zimmermann, A., and De Raedt, L.** Cluster-grouping: From subgroup discovery to clustering. *Machine Learning (ML) 77* (2009), 125–159.

[214] **Zucchini, W., MacDonald, I. L., and Langrock, R.** *Hidden Markov models for time series: An introduction using R*, 2nd ed. Chapman and Hall/CRC: Boca Raton, 2017.

Summary

In this dissertation, we analyze variation in human behavior using a Local Pattern Mining (LPM) framework called Exceptional Model Mining (EMM). EMM aims to discover subgroups in a population that somehow behave exceptionally. These subgroups are described using an interpretable language of conjunctions of attribute-value pairs.

We observe that employing EMM in real-world use cases is challenging, since data is often hierarchically structured. Although formal definitions in EMM are agnostic about the origin of the data and whether or not one observation is independent from the next one, most existing EMM methodologies assume data to be in the conventional data mining representation where each individual can be described with one tuple of attribute-values. In contrast, in hierarchical data, attributes may contain multiple values per individual and individuals may be described by tuples of tuples. Then, the EMM framework runs into problems with the selection and description of candidate subgroups and with assessing exceptionality. Therefore, the main research question in this dissertation is:

How to discover exceptional subgroups in hierarchical data?

To answer this question, we first formally define hierarchical data as a collection of measurements taken from various types of entities, where the measurements of one entity type are nested in the entities of another entity type. Second, we propose a unified terminology that classifies existing EMM methodologies based on whether descriptors and targets reside at lower, the same or higher hierarchical levels than the subgroup level, the hierarchical level of the entity type for which subgroups should be formed.

The work presented in this dissertation significantly contributes to the body of scientific literature on EMM for hierarchical data. Our work further populates rather sparse categories in our unified terminology. Specifically, we develop new EMM methodologies for three types of hierarchical data:

1. we analyze sequential data in target space by employing Markov chains of varying order and develop three new quality measures based on information-theoretic scoring functions,
2. we develop EMM-RCS, an EMM instance for Repeated Cross-Sectional data that deals with practical RCS data problems, including uneven spacing of measurements over time, fluctuating sample sizes and missing data,
3. we consider hierarchal data with nested observations in both descriptive and target space and propose the concept of aggregated descriptors as a generic approach to flattening nested data in descriptive space.

In addition, throughout this dissertation, we demonstrate external validity of our proposed EMM methodologies by enabling domain experts to confirm existing hypotheses and to spawn interest for new theories. In *diabetes care*, domain experts confirm that our findings support the hypothesis that monitoring blood glucose values using an iCGM device is a valuable new way to monitor glycemic treatment for patients with diabetes type 2. In *public health*, domain experts interpreted our findings and conclude that they confirm existing knowledge, spawn two hypotheses for further sociological research and provide disconfirming evidence to a sociological hypothesis that aligns with existing studies. In *learning analytics*, based on our findings, domain experts discovered subgroups of children with dyscalculia. Our findings furthermore confirm the belief that numerical processing competences strongly correlate with arithmetic skills and reveal promising opportunities for further development of digital assessment tools.

In sum, in this dissertation, we advance the framework of Exceptional Model Mining by proposing generic and domain-independent methodologies for discovering exceptional subgroups in various types of hierarchical data. Given the potential societal impact of the patterns discovered with our proposed methods, we believe integration of EMM methods into real-world applications and software is within the realm of possibility. Further research in this direction would be valuable and the work presented in this dissertation could very well serve as a starting point.

Curriculum Vitæ

Rianne Margaretha Schouten was born on May 11th 1991 in Goor, the Netherlands. She completed a Bachelor's degree in Medicine in 2012 and a Master's degree in Methodology and Statistics for the Behavioral, Biomedical and Social Sciences in 2017, both at Utrecht University, the Netherlands. During her studies, Rianne contributed to state-of-the-art research in the medical domain, and in space research, by supporting domain experts in doing statistically sound and trustworthy analyses. In 2016, she started her Master's thesis on the topic of *generating missing values in complete data* and visited the STAN development team of Professor Andrew Gelman at Columbia University in the City of New York. Her work resulted in two publications and an open source software package called `ampute`. After her studies, Rianne continued doing research on evaluating missing data methods. Meanwhile, she worked as a data scientist (1 year) and data warehouse developer (2 years).

In 2020, Rianne started her PhD program in the Data Mining Group at TU/e under the supervision of prof. dr. Mykola Pechenizkiy and dr. Wouter Duivesteijn. Her PhD thesis on the topic of *Exceptional Model Mining for Hierarchical Data* contains both methodological advances and domain-specific contributions. Overall, Rianne contributed to 16 papers, of which she was the first author for nine. Furthermore, she loves working in inter- and multidisciplinary teams and set up multiple fruitful collaborations that resulted in joint publications at top-level conferences (DAMI, SDM, ECML PKDD). In addition, she lectured 4 Master's courses and supervised ~15 groups of students during research projects. She contributed to the supervision of ~10 individual Master students. Several of these projects have resulted in publications at conferences such as IDA, DS and BNAIC/BeNeLearn. In 2021 and in 2022, Rianne received an award for excellent teaching evaluations.

Rianne is an active member of the research community. She served on program committees for top AI conferences and journals, including ECML PKDD, DAMI, ML and the Journal of the Royal Statistical Society. In 2024, Rianne was recognized as an excellent reviewer at ECML PKDD. In the same conference, she additionally served as Proceedings Chair. Furthermore, Rianne contributed to obtaining two funding packages for TU/e. First, in 2022, she contributed to receiving a seed money grant of €45k on the topic of multiple imputation using GANs. And recently, in 2024, Rianne obtained a personal grant of €40k to investigate the feasibility of integrating local pattern mining techniques into digital learning platforms. She aims to continue working on this relevant and important topic as a post-doc at TU/e.

Acknowledgments

A long list of amazing people have contributed to the creation of this dissertation. Honestly, kind words do not do justice to what some people have done for me. But I hope that the upcoming paragraphs reflect some of my heartfelt gratitude.

Obviously, the first in line is prof. dr. Mykola Pechenizkiy. From the moment I met Mykola, I appreciate how he combines a calm and friendly communication style with being thorough and to-the-point on the content. During my PhD program, I have always felt that Mykola was able to emphasize my strong characteristics, without disregarding the weaker ones. Moreover, he would actively point me towards interesting opportunities and additionally provide emotional support when needed. Mykola, you played a crucial role in my professional development. Thank you for all you have done for me.

A close runner-up is dr. Wouter Duivesteijn, whose passionate and dedicated communication style created an inspiring working environment. Wouter, I greatly value your initiative to lunch together, to organize EMM-lab, to play poker and to always talk in English. More than that, I appreciate the way you listened to me and how you spend a tremendous amount of time and effort to prepare for our meetings and to follow up on open questions. I felt treated as an equal. Thank you for investing in me and in us; it has brought me much more than the nice list of scientific articles we wrote together.

I would like to thank the committee members of my PhD defense for assessing and approving my dissertation and for providing valuable feedback: prof. dr. Martin Atzmueller (Osnabrück University), prof. dr. Barbara Hammer (Bielefeld University), prof. dr. Claudia Plant (University of Vienna), prof. dr. Bruno Crémilleux (University of Caen Normandy) and prof. dr. Anna Vilanova (TU/e). I would also like to express my gratitude to prof. dr. Hermie Hermens (UTwente), who initiated the consortium that lead to the Commit2Data project in which I was hired and where we developed the components of an Exceptional and Deep Intelligent Coach (EDIC). Prof. dr. Hermie Hermens additionally provided me with valuable advice during an emotional phase in my life. Thank you Hermie.

A lot of the work presented in this dissertation is the result of an intensive collaboration with domain experts from across the globe. These collaborators deserve special acknowledgment for their willingness to spend time and effort in building bridges. Many thanks to: prof. dr. Goos Laverman (ZGT hospital), dr. Niala den Braber (ZGT hospital), dr. Gonneke W.J.M. Stevens (Utrecht University), dr. Saskia A.F.M. van Dorsselear (National Institute of Mental Health and Addiction), dr. Elisa L. Duinhof (National Institute of Mental Health and Addiction), dr. Karin Monshouwer (National Institute of Mental Health and Addiction), dr. Jolanda Luime (Maxima Medical Center), dr. Jacob M. Paul (University of Melbourne), dr. Jose M. Luna (University of Cordoba) and prof. dr. Pekka Räasänen (Turku Research Institute for Learning Analytics). Working with you has been motivating and rewarding and I am proud of the results we obtained together.

Furthermore, thank you dr. Zahra Atashgahi, dr. Niala den Braber, Carlijn Braem and Ellen Kuipers for many pleasant interactions during and outside EDIC progress meetings.

During the PhD program, I have enjoyed being part of the Data & AI cluster at TU/e, where so many friendly colleagues are so committed to contributing to AI development. I cannot count the number of times that I requested help and received instant replies, suggestions and advice from various directions. Thank you all. I would like to particularly acknowledge valuable input I received from dr. Marcos L.P. Bueno, dr. Robert Peharz, Hilde Weerts, Simon Koop, dr. Pieter Gijsberts, dr. Sibylle Hess, prof. dr. George Fletcher and prof. dr. Cassio de Campos. Thank you Julie and Gizem, for all the work you do in the background, for organizing great activities and for distributing plenty of smiles and encouragements. Thanks to my paranymphs: Bram Grooten, who is one of the kindest persons I have ever met, and Ellen van de Looij-Guns, who is always interested and compassionate.

Before joining TU/e, I have had the pleasure to work with some very inspiring people. In particular, I would like to mention my previous manager Kees Snoek, who shared many insights and stories with me. Kees, I learned a lot from you. Thanks as well to Nelleke Shioda, who taught me the importance of building personal relationships with colleagues. Furthermore, I would like to acknowledge dr. Gerko Vink (Utrecht University), whose enthusiasm is contagious! Thank you Gerko, for supporting me in the early phases of my career. Looking back, I realize I have not always recognized the opportunities you created for me. Thank you for not letting me down nonetheless. I am especially grateful that you arranged that I could join you at Columbia University. More than that, you adopted me into your family. I will never forget our trip to Long Island and the amazing hugs I received from Joris and Guus. Thank you Aly and thank you Gerko.

Some of my closest friends are my greatest examples. Dr. Rinske Ijsselhof, thank you for giving me the opportunity to work with you. Marieke Hagg, I remember our coffee breaks and you explaining to me what a PhD is. Dr. Tessa Goes-Roelofs, you were the first to demonstrate to me that obtaining a PhD degree can be done. You are the first in many things and I am grateful that I can be a part of it. Dr. Elisa ter Harmsel-Duinhof, thank you for letting me be your paranymph. It was special for me to be close to you in that moment.

As with anything in life, a PhD trajectory comes with ups and downs. In my case, there was a phase where I had completely lost my confidence. I doubted my skills as a mother, and my skills in general. I am deeply grateful to the amazing people from De Boeiende Kunst (an art studio in Eindhoven) and to Marissa Rovers (an expert in supporting young mothers), who helped me get back on my feet.

Papa en mama, jullie zijn geweldig. Dank je wel voor het voorbeeld dat jullie voor me zijn. Zo ken ik weinig mensen die zo’n groot leervermogen hebben als jullie. Ook zijn jullie harde werkers en hebben jullie een enorm doorzettingsvermogen. In momenten dat alles om me heen lijkt weg te vallen, zijn jullie daar. Ik hoop dat ik voor Arya kan zijn zoals jullie voor mij zijn.

Het is fijn om familie te hebben. Dank je wel dr. Jop Schouten, voor je gezelligheid. Ik herken veel van mezelf in jou. Dank je wel Anouk, dat je er bent, en dat Arya lekker met

Milan en Jelte mag spelen. Dank je wel Maaike, voor je rust en je krachtige manier van communiceren. De voorkant van deze dissertatie is heel mooi geworden en ik ben vooral verliefd op het figuur dat je gemaakt hebt voor bladzijde vii. Lieve Moos, ik lijk zoveel op jou dat het soms een beetje confronterend is. Bedankt dat je dan vergevingsgezind bent. Ik ben trots op hoe jij je weg vindt en weet zeker dat je je doelen gaat bereiken. Bedankt Ika, voor wie je bent voor Moos. Lieve allemaal, dank jullie wel voor jullie hulp en steun op verwachte en onverwachte momenten.

Dear Raghu, you never doubted me, neither my motherhood skills nor my ability to succeed in obtaining a PhD degree. You supported me with many selfless sacrifices, without complaining and with full commitment. I am still wondering how the universe brought us together, but it was the best thing that happened to me. Jij bent de allersterkste papa. Raghu, a big kiss from me to you.

Arya, you are exceptional.

Rianne Margaretha Schouten
Eindhoven, December 2024

SIKS dissertations

2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UvA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web

19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UvA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UvA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UvA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TU/e), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UvA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VUA), Logics for causal inference under uncertainty

23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VUA), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VUA), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TU/e), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TU/e), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TU/e), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech

30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

---

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data

06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OU), Learning visually grounded and multilingual representations

---

2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TU/e), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OU), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models

12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst

31 Gongjin Lan (VUA), Learning better – From Baby to Better

32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

---

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (UL), Robust rules for prediction and description

24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

---

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications

12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries

13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation

14 Selma Čaušević (TUD), Energy resilience through self-organization

15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight

18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation

19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals

20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning

21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain

22 Alireza Shojaifar (UU), Volitional Cybersecurity

23 Theo Theunissen (UU), Documentation in Continuous Software Development

24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning

25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs

26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour

27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions

28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts

29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education

02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis

04 Mike Huisman (UL), Understanding Deep Meta-Learning

05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair

06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence

07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems

08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation

09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks

10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science

11 withdrawn

12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning

13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence

14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling

15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health

16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems

17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design

18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations

19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior

20 Ritsart Anne Plantenga (UL), Omgang met Regels

21 Federica Vinella (UU), Crowdsourcing User-Centered Teams

22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts

23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution

24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour

25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics

26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning

28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases

30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting

31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows

32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining

33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction

34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method

35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach

36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs

37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models

38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings

39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance

40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models

41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines

42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis

43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms

44 Marcel Schmitz (OU), Towards learning analytics-supported learning design

45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making

46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology

47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs

48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems

02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks

03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI

04 **Rianne Schouten** (TU/e), Exceptional Model Mining for Hierarchical Data

ERRATUM

PhD thesis: Exceptional Model Mining for Hierarchical Data

Rianne M. Schouten, 14 January 2025

In Chapter 7, we discuss the effect of several anti-redundancy and validation techniques on the set of discovered subgroups. Tables 7.2 - 7.4 contain the results *before* pruning. However, in the text, the numbering of subgroups refers to results *after* pruning. Below, we present a list of corrections.

On page 113, in Figure 7.1b, the *green* line listed as **5** should be colored with ***dark purple*** and be listed as ***6***, in accordance with the **6** in the bottom row in Table 7.3.

On page 114, par. 2, *subgroups 11 and 15* should be changed to *subgroups 15 and 20.*

On page 114, par. 2, *subgroups 1, 4, 5 and 8* should be changed to *subgroups 1, 2, 3, 4, 7 and 8.*

On page 114, par. 2, *subgroups 11 and 15* should be changed to *subgroups 15 and 20.*

On page 114, par. 3, *trend groups 2 and 5 (purple and green lines)* should be changed to *trend groups 2 and 6 (purple and dark purple lines).*

On page 115, par. 1, *(subgroups 3 and 9)* should be changed to *(subgroups 6 and 14).*

On page 115, par. 1, *(subgroup 12)* should be changed to *(subgroup 18).*

On page 115, par. 2, *trend group 5 (subgroup 11)* should be changed to *trend group 6 (subgroup 16).*

On page 115, par. 2, *(subgroup 6)* should be changed to *(subgroup 10).*

On page 115, par. 2, *at least moderately urbanized* should be changed to *at most moderately urbanized.*

On page 116, par. 2, *(trend groups 2 and 5)* should be changed to *(trend groups 2 and 6).*

On page 117, par. 2, *subgroups 2, 6, 9, 10* should be changed to *subgroups 2, 6, 15, 20.*

On page 117, par. 2, *subgroups 5, 8* should be changed to *subgroups 5, 11.*

On page 117, par. 5, *subgroup 11 in Table 7.3* should be changed to *subgroup 16 in Table 7.3.*

On page 117, par. 5, *subgroups 6, 7 in Table 7.4* should be changed to *subgroups 6, 9 in Table 7.4.*

On page 118, par. 1, *subgroups 2, 9 in Table 7.4* should be changed to *subgroups 2, 15 in Table 7.4.*

On page 118, par. 1, *subgroups 2, 7, 13 in Table 7.3* should be changed to *subgroups 5, 11, 19.*