Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS)

Rianne Margaretha Schouten[*] Wouter Duivesteijn[*] Mykola Pechenizkiy[*]

Abstract

Repeated Cross-Sectional (RCS) data measures a phenomenon by repeatedly sampling new cases 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.

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 [19, 23] (see black line in Figure 3a in Section 7.1). Since adolescent alcohol consumption has short-term risks (e.g., injuries, violence) and long-term risk of adult alcohol dependence [14], 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% [6]. 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 [10, 14]. 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 paper, we develop EMM-RCS: a new instance of the framework of Exceptional Model Mining (EMM) [5, 11] that seeks subgroups with exceptional societal trends in Repeated Cross-Sectional data.

EMM is a local pattern mining framework 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. Another set of attributes is used to describe subgroups as a conjunction of attribute-value conditions. 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 [2]. RCS data is obtained by sampling new cases from a population at successive occasions. It differs from time series where multiple measurements are taken per case with very short time intervals, and from longitudinal data where the same people are followed through time (see Figure 1). Instead, RCS research collects the same information from different cases 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).

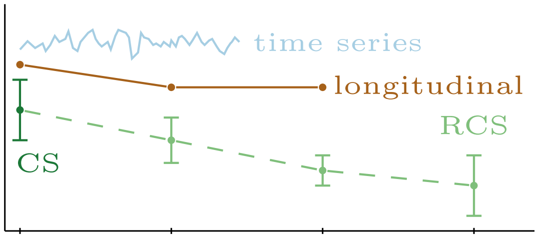


Figure 1: Schematic overview of various types of data. In time series, multiple measurements are sampled per case with very short time intervals. In longitudinal data, sampling is done with long intervals in a relatively long period of time. Repeated cross-sectional (RCS) data is collected from new samples at each measurement occasion, resulting in varying sample sizes.

[*] Eindhoven University of Technology, the Netherlands, {r.m.schouten,w.duivesteijn,m.pechenizkiy}@tue.nl

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 (e.g. [17]) but as in time series or longitudinal data, there the sequence is known per case and the sample size is fixed. Instead, in RCS data, a case 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 [3]. 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.

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 [16], British Social Attitudes (cf. `https://bsa.natcen.ac.uk/`), and Monitoring the Future [9]. 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 [14]. 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) [5, 11] 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 [20] and an EMM model class for A/B tests [4]. Note that EMM model classes exist for sequential data [15, 21]. Similarly, methods exist to detect time series anomalies or discords [13]. However, in both sequential and time series data the repeated measurements are taken within cases (cf. Figure 1), which requires different methods than analyzing change over time in RCS data (where cases 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) [7] with a numerical target [12], 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 [18] 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, which is related to but not the same as standard error.

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 cases [2]. 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) [23] collects data with 4-year intervals. In Section 7.1, we use its 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. The former makes trend analysis easier.

The 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 [1]. An estimator performs well if it produces unbiased and precise estimates and its performance depends largely on the sampling design [1]. In this paper, 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 case 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.

3.1 Exceptional Model Mining

Exceptional Model Mining (EMM) [5, 11] seeks subgroups in a dataset that somehow behave exceptionally. Here, the dataset Ω is defined as a bag of N records $r \in \Omega$ of the form $r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m)$; we distinguish k descriptive attributes and m target attributes. The former are used to form interpretable subgroups by deploying a rule-based description language using conjunctions of attribute-value conditions. For instance, a subgroup of adolescents can be described as *11* $\leq$ *age* $\leq$ *15* $\wedge$ *school year = 4* $\wedge$ *lives with both parents = yes* (cf. Section 7.1).

The choice of model over the target attributes is called the *model class*. Next, a quality measure quantifies how different the model in the subgroup is from the model in a reference group (often the entire dataset). EMM searches through the space of possible subgroups and outputs the top-q most exceptional subgroups it encounters.

4 Exceptional Model Mining for Repeated Cross-Sectional Data (EMM-RCS)

RCS data does not follow the format of Section 3.1. We therefore redefine our notion of data as follows:

Definition 4.1. (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 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; $n_{x_t} \neq n_{x_{t'}}$ $(t \neq t')$.

Definition 4.1 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 $x_t \in \mathcal{T}$.

4.1 Descriptive space

Denoting the collective domain of the descriptive attributes $(a_1, a_2, \ldots, a_k)$ by $\mathcal{A}$, a *description* formally is a function $D : \mathcal{A} \mapsto \{0, 1\}$. A record $r^i_{x_t}$ is *covered* by D if and only if $D(a^i_1, a^i_2, \ldots, a^i_k) = 1$. We define a subgroup as follows:

Definition 4.2. (Subgroup) *The subgroup corresponding to a description* D *is the bag of records* $SG_D \subseteq \Psi$ *that* D *covers:*

$$SG_D = \{r^i_{x_t} \in \Omega_{x_t} \mid D(a^i_1, \ldots, a^i_k) = 1, \Omega_{x_t} \in \Psi\}.$$

A description D thus collects records 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 [3]. 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 record $r^i_{x_t}$; whether or not record $r^i_{x_t}$ 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 $i \in n_{x_t}$ records $r^i_{x_t}$ at occasion x_t. The second type makes that value a_j could be missing for a specific record $r^i_{x_t}$, but be observed for another record $r^{i'}_{x_t}$ 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 (cf. Section 3.1) 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. [5, Section 4.1]). We add a condition where the attribute-value pair is missing.

Definition 4.3. (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 $\eta(D)$.*

See [22, Section 5] for a discussion why popular missing data methods such as dropping incomplete cases or missing value imputation are not sufficient for RCS data, and a demonstration of the working of Definition 4.3 in an experiment.

Refinement strategies exist for binary, numerical, and nominal descriptive attributes [5]. In RCS data, distinguishing nominal from ordinal attributes is practically relevant. Hence, we define a refinement strategy for ordinal attributes.

Definition 4.4. (Refining ordinal attribute) *For an ordinal attribute a_j, order the unique values of a_j; this gives a list of ordered values $w_1, \dots, w_m$. Then, add $\{D \cap (a_j \leq w_h), D \cap (a_j > w_h)\}_{h=1}^{m-1}$ to the set of descriptions $\eta(D)$.*

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 $\mathcal{D}$, 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.

Definition 4.5. (Quality measure) *A quality measure is a function $\varphi : \mathcal{D} \mapsto \mathbb{R}$ that assigns a numeric value to a description D.*

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

$$\varphi_{RCS}(D) = f\left(\{z_{x_t} \mid x_t \in \mathcal{T}\}\right) \tag{4.1}$$

$$z_{x_t} = \frac{\left|\theta^{SG}_{x_t} - \theta^0_{x_t}\right|}{se\left(\theta^{SG}_{x_t}\right)}. \tag{4.2}$$

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 (4.1) 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 5.2 and now focus on Equation (4.2).

In Equation (4.2), $\theta^{SG}_{x_t}$ 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 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.

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}.

5 Instantiations of our quality measure

Before we apply our method to both synthetic and real-world data in Sections 6 and 7, we will now first give examples of choices for μ, θ, and f.

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 [6]. To seek exceptional trends in alcohol use (cf. Section 7.1), 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 [1]. Then, μ_{x_t} can be approximated with the proportion of the sampled values ℓ, and corresponding standard error

$$\mu_{x_t} = \frac{1}{n_{x_t}} \sum_{i=1}^{n_{x_t}} \ell^i_{x_t} \tag{5.3}$$

$$se\left(\mu_{x_t}\right) = \sqrt{\frac{\mu_{x_t}\left(1-\mu_{x_t}\right)}{n_{x_t}-1}}. \tag{5.4}$$

While analyzing the Eurobarometer dataset [22, Section 4], 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 (5.3) with standard error

$$se\left(\mu_{x_t}\right) = \frac{sd\left(\mu_{x_t}\right)}{\sqrt{n_{x_t}}} = \left(\frac{\sum_{i=1}^{n_{x_t}}\left(\ell^i_{x_t}-\mu_{x_t}\right)^2}{\sqrt{n_{x_t}}\left(n_{x_t}-1\right)}\right). \tag{5.5}$$

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 (5.3) the slope and its standard error are

$$\theta_{x_t} = \mu_{x_{t+1}} - \mu_{x_t} \qquad \forall t \in \{1,2,\ldots,T-1\} \tag{5.6}$$

$$se\left(\theta_{x_t}\right) = \sqrt{se\left(\mu_{x_{t+1}}\right)^2 + se\left(\mu_{x_t}\right)^2}. \tag{5.7}$$

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,

$$\tau_{x_t} = \frac{\sum_{t=1}^{u} w^*_{x_t}\mu_{x_t}}{\sum_{t=1}^{u} w^*_{x_t}} = \sum_{t=1}^{u} w_{x_t}\mu_{x_t}, \tag{5.8}$$

where $w_{x_t} = {w^*_{x_t}}/{\sum_{t=1}^{u} w^*_{x_t}}$. In Section 7.1, 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 is

$$se\left(\tau_{x_t}\right) = \sqrt{\sum_{t=1}^{u} w^2_{x_t} se\left(\mu_{x_t}\right)^2}. \tag{5.9}$$

As a second step, we can then define θ_{x_t} and its standard error as in Equations (5.6) and (5.7), 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 (4.1) is $T-2$.

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 $x_t \in \mathcal{T}$. 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 [22, Section 4], 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.

In Section 7.1, the general trend in alcohol use is predominantly decreasing (see Figure 3a). 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 (4.2) decreases if $se(\theta^{SG}_{x_t})$ increases.

We experiment with two solutions. First, we do not correct for varying sample sizes ($se(\theta^{SG}_{x_t}) = 1$) 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}). A comparison between these two approaches and a discussion of how the threshold value can be set is given in [22, Section 3].

6 Synthetic data experiments

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 $N = 10\,000$ cases and randomly assign cases to one out of $T = 10$ measurement occasions. Second, we draw $ncovs = 10$ binary descriptors $a_1, \ldots, a_{10}$, each from a binomial distribution $\text{Bin}\,(n = N, p = 0.5)$. 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. For instance, a subgroup with 2 literals could be described by $a_4 = 1 \wedge a_7 = 1$. 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 cases. For these cases, 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 (5.3) and (5.5), $\theta^0 = \theta^\Psi$ and f_{max}. Throughout this paper, we search the space of candidate subgroups using beam search. See [22, Section 1] for more on beam search, the choice of search parameters and applied anti-redundancy and validation techniques. Here, $q = 20$, $d = 5$, and $w = 20$. Every combination of simulation conditions is repeated $nreps = 100$ times. All simulation code can be found in our Github repository at `https://github.com/RianneSchouten/EMM_RCS`.

Figure 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 (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).

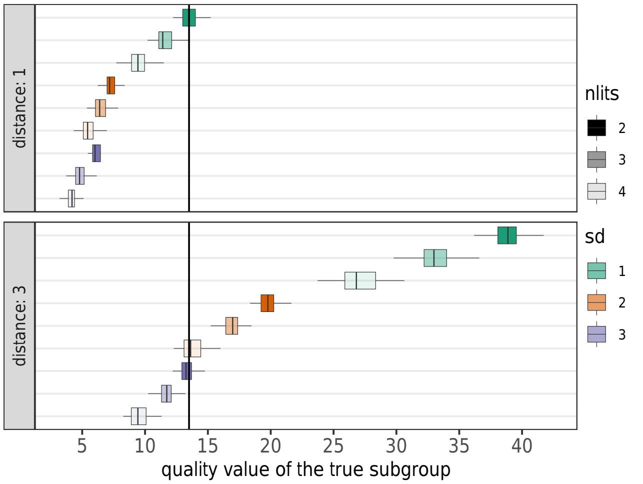


Figure 2: Boxplots of the quality values of the ground truth subgroup for 100 repetitions. Top and bottom panel mark the distance between the subgroup and the non-subgroup, $sd \in \{1, 2, 3\}$ specifies the standard deviation of the trend in the subgroup and $nlits \in \{2, 3, 4\}$ specifies the number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

Figure 2 furthermore displays the tradeoff 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 with a trend line at $dist = 1$ with $sd = 1$ and $nlits = 2$ (dark green boxplot at the top) has the same quality value as a subgroup with $dist = 3$, $sd = 3$ and $nlits = 2$ (dark purple boxplot in bottom panel, see vertical line). 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.

More in-depth analysis of results on synthetic data can be found in [22, Section 2]. Across 27 parameterizations (some adversarial towards EMM-RCS) the ground truth subgroup achieved median rank 1 in 20 cases.

7 Real-world data experiments

We run experiments on three real-world datasets, using various combinations of θ and f. The results for two datasets are discussed below, and the third experiment can be found in [22, Section 4].

7.1 HBSC and DNSSSU dataset Alcohol use among Dutch adolescents is monitored by two studies: the Health Behaviour in School-aged Children study (HBSC) [23] and the Dutch National School Survey on Substance Use (DNSSSU) [19]. The two studies are conducted every 4 years in alternating fashion, resulting in a nice regular time interval of 2 years when combining the two datasets. We analyze the trend in alcohol use between 2003 and 2019; encompassing the HBSC data from 2005, 2009, 2013, and 2017, and the DNSSSU data from 2003, 2007, 2011, 2015, and 2019. The dataset contains 36 306 cases.

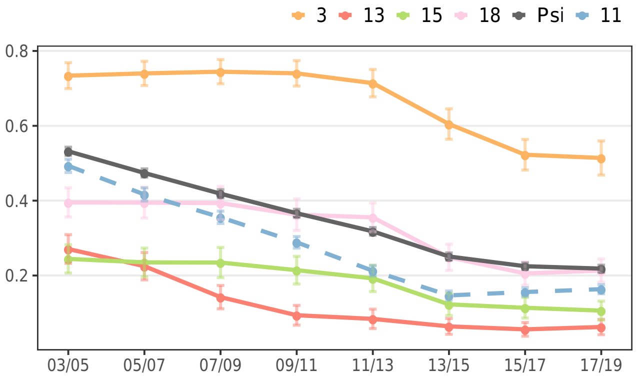
We investigate trends in alcohol use in the last 4 weeks by using a binary attribute with a binomial distribution (cf. Section 5.1). Specifically, we calculate a weighted moving average of the prevalence (cf. Equations (5.3) and (5.8)). In consultation with domain experts the dataset is constructed such that we have 10 descriptive attributes without missing values. Two of them are binary, 2 numerical, 3 nominal, and 3 ordinal.

Figure 3 displays trends in alcohol use among Dutch adolescents across the population (black) and of some exceptional subgroups. Figure 3b gives the coverage and descriptions of the subgroups shown in Figure 3a.

The subgroups with a solid trend line are found by comparing the slope of the moving average of two subsequent prevalence estimates against a reference value of $\theta^0_{x_t} = 0$ and by subsequently counting the number of z-scores that fall within the threshold value $\epsilon = 0.01$. As can be seen in Figure 3a, we find several subgroups with horizontal parts in their trend. As expected, we find subgroups of adolescents that drink more (#3), similar (#18) or less (#13, #15) than the average adolescent. The descriptions give domain experts valuable knowledge about whom to target with campaigns. For instance, it would be smart to target adolescents in the fourth year of secondary school (#3) rather than adolescents in the first two years (#13).

By counting the number of horizontal slopes, our quality measure is not restricted to fixed measurement occasions. We clearly see this in Figure 3a, where the trend is horizontal between 2003 and 2013 for subgroups 3, 15, and 18, while alcohol use in subgroup 13 decreases in that same period. It would be interesting to further explore the relation between these groups of adolescents and government campaigns that ran in those years.

The dashed line represents a subgroup that we find by comparing the slope of the moving average with the overall trend in the population. We found no subgroups with an increase in alcohol use at any measurement occasion. We did find subgroups with flat parts in their trend (such as the ones displayed in Figure 3a but with different descriptions) and subgroups with a stronger decrease in alcohol use, such as subgroup 11, which covers younger adolescents in higher educational tracks.



(a) Trends of subgroups, and the overall population (black).

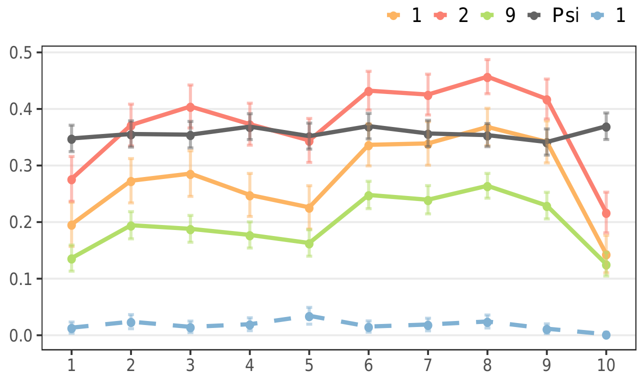
#	Cov.	Description
3	0.07	$11 \leq$ age $\leq 15 \wedge$ school year $= 4 \wedge$ lives with both parents = {yes}
13	0.06	life satisfaction $= 10 \wedge$ school year = {1,2} $\wedge$ does mother have a job = {yes, no}
15	0.06	ethnic group = {non-Western} $\wedge$ school level $\neq$ {VWO} $\wedge$ sex = {boy}
18	0.07	ethnic group = {non-Dutch} $\wedge$ $13 \leq$ age $\leq 19 \wedge$ life satisfaction ≤ 7
11	0.41	school level = {HAVO, VWO} $\wedge$ $10 \leq$ age ≤ 15

(b) Coverage and description of the subgroups.

Figure 3: Subgroups displaying unusual moving average of the prevalence of alcohol use among Dutch adolescents. Subgroups with solid trend lines are found by counting how many slopes (of two subsequent measurement occasions) are smaller than $\epsilon = 0.01$ (cf. Section 5.2). The subgroup with a dashed trend line is found by comparing the slope in the subgroup with the slope in the overall population ($\theta^0 = \theta^\Psi$).

7.2 Brexit dataset A 10-wave survey examines Attitudes Towards Brexit (ATB) in the aftermath of the 2016 Brexit referendum on EU membership. 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 [8]. The data is available in the UK Data Service, at `https://reshare.ukdataservice.ac.uk/854869/`.

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*).



(a) Trends of subgroups, and the overall population (black).

#	Cov.	Description
1	0.32	govthand = {don't know, very, fairly badly} ∧ tradeimmig ≤ 7 ∧ age ≥ 47
2	0.4	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

(b) Coverage and description of the subgroups.

Figure 4: Subgroups displaying unusual trends of the proportion of people who think of themselves as a *leaver* considering Brexit. The x-axis represents 10 study waves between April 25, 2017 and January 10, 2020. The subgroup with a dashed trend line is found by directly comparing the proportion in the subgroup with the proportion in the overall population ($\theta_{x_t} = \mu_{x_t}$). Subgroups with solid trend lines are found by comparing the slopes of the proportion ($\theta_{x_t} = \mu_{x_{t+1}} - \mu_{x_t}$).

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 cases.

In the Brexit dataset, we explore trends of the proportion of people who think of themselves as a *leaver*. Results can be found in Figure 4. The population trend is fairly horizontal, with an approximate average of 35% of respondents who want to leave the European Union.

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. Figure 4b). Most of the top-20 subgroups revolve around *hindsight* = *wrong*, and we do not find subgroups that cover more leavers (than remainers) or subgroups where the trend is not horizontal. This does not mean that those subgroups do not exist. Rather, by using Equation (5.3) and $f_{\max}$, EMM-RCS finds subgroups with a maximal difference at just one measurement occasion, and apparently there are no subgroups with a deviation larger than 0.35.

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 Figure 4b 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.

8 Conclusion

We propose Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS): a method finding subgroups with exceptional trends in data collected with a Repeated Cross-Sectional (RCS) 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.

Perhaps the starkest illustration of the versatility of EMM-RCS and our quality measure is provided by the results in Figure 4, 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.

When analyzing trends among Dutch adolescents, we find a subgroup whose alcohol use has not been influenced by government campaigns. In future work, we intend to explore both global and local trend exceptionalities on this dataset together with the domain experts, in a bid to more precisely target government campaigns aimed at reducing adolescent alcohol consumption.

Acknowledgments

We gratefully acknowledge all people who study societal trends and thereby contribute to policy making that benefits all of us. In particular, we thank Elisa Duinhof, Gonneke Stevens, Karin Monshouwer and Saskia Dorsselaer for providing domain expertise on the topic of alcohol use among Dutch adolescents. We furthermore acknowledge the quick and constructive help of Koo Rijpkema and Peter Lugtig. This work is part of the research program Data2People with project EDIC and partly financed by the Dutch Research Council (NWO).

References

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

[2] A. Bryman. *Social research methods*. Oxford University Press, 2016.

[3] CBS. Jaarrapport 2016 landelijke jeugdmonitor. Den Haag/Heerlen: CBS, 2016.

[4] W. Duivesteijn, T. Farzami, T. Putman, E. Peer, et al. Have it both ways — from A/B testing to A&B testing with exceptional model mining. In *Proc. ECMLPKDD*, pages 114–126, 2017.

[5] W. Duivesteijn, A. J. Feelders, and A. Knobbe. Exceptional model mining. *Data Mining and Knowledge Discovery*, 30(1):47–98, 2016.

[6] A. van Giessen, J. Boer, I. van Gestel, E. Douma, et al. *Voortgangsrapportage Nationaal Preventieakkoord 2019*. RIVM, 2020.

[7] F. Herrera, C. J. Carmona, P. González, and M. J. Del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3):495–525, 2011.

[8] S. B. Hobolt, T. J. Leeper, and J. Tilley. Divided by the vote: Affective polarization in the wake of the Brexit referendum. *British Journal of Political Science*, pages 1–18, 2020.

[9] L. D. Johnston, R. A. Miech, P. M. O'Malley, J. G. Bachman, et al. Monitoring the future national survey results on drug use, 1975-2020: Overview, key findings on adolescent drug use. *Institute for Social Research*, 2021.

[10] M. R. Kern, E. L. Duinhof, S. D. Walsh, A. Cosma, et al. Intersectionality and adolescent mental wellbeing: a cross-nationally comparative analysis of the interplay between immigration background, socioeconomic status and gender. *Journal of Adolescent Health*, 66(6):S12–S20, 2020.

[11] D. Leman, A. Feelders, and A. Knobbe. Exceptional model mining. In *Proc. ECMLPKDD*, pages 1–16, 2008.

[12] F. Lemmerich, M. Atzmueller, and F. Puppe. Fast exhaustive subgroup discovery with numerical target concepts. *Data Mining and Knowledge Discovery*, 30(3):711–762, 2016.

[13] X. Li and J. Han. Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proc. VLDB*, pages 447–458, 2007.

[14] M. E. de Looze, S. A. F. M. van Dorsselaer, K. Monshouwer, and W. A. M. Vollebergh. 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*, 50:90–101, 2017.

[15] R. Mathonat, D. Nurbakova, J.-F. Boulicaut, and M. Kaytoue. Anytime mining of sequential discriminative patterns in labeled sequences. *Knowledge and Information Systems*, 63(2):439–476, 2021.

[16] S. Mokinaro, J. Vincente, E. Benedetti, S. Cerrai, et al. ESPAD report 2019: Results from European School Survey Project on Alcohol and other Drugs. *Technological University Dublin*, 2020.

[17] D. Mollenhauer and M. Atzmueller. Sequential Exceptional Pattern Discovery using Pattern-Growth: An extensible framework for interpretable machine learning on sequential data. In *Proc. XI-ML*, 2020.

[18] B. F. I. Pieters, A. Knobbe, and S. Dzeroski. Subgroup discovery in ranked data, with an application to gene set enrichment. In *Proc. PL*, pages 1–18, 2010.

[19] M. Rombouts, S. A. F. M. van Dorsselaer, T. Scheffers-van Schayck, M. Tuithof, et al. *Jeugd en riskant gedrag 2019. Kerngegevens uit het Peilstationsonderzoek Scholieren*. Trimbos-Instituut, Utrecht, 2020.

[20] K. Rudaś and S. Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, 32(5):1275–1305, 2018.

[21] R. M. Schouten, M. L. P. Bueno, W. Duivesteijn, and M. Pechenizkiy. Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. *Data Mining and Knowledge Discovery*, 2021.

[22] R. M. Schouten, W. Duivesteijn, and M. Pechenizkiy. Exceptional model mining for repeated cross-sectional data (EMM-RCS) — supplementary material. Technical report, available at Figshare, `https://doi.org/10.6084/m9.figshare.18688220`, 2022.

[23] G. W. J. M. Stevens, S. A. F. M. van Dorsselaer, M. Boer, S. de Roos, et al. *HBSC 2017. Gezondheid en welzijn van jongeren in Nederland*. Utrecht University, 2018.