Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining: Discovering exceptional trends in adolescent alcohol use in the Netherlands

Rianne M. Schouten[1] (✉), Gonneke W.J.M. Stevens[2], Saskia A.F.M. van Dorsselaer[3], Elisa L. Duinhof[3], Karin Monshouwer[3], Mykola Pechenizkiy[1], and Wouter Duivesteijn[1]

[1] Eindhoven University of Technology, Data Mining Group, Eindhoven, the Netherlands, {r.m.schouten,m.pechenizkiy,w.duivesteijn}@tue.nl
[2] Utrecht University, Department of Interdisciplinary Social Science, Utrecht, the Netherlands, g.w.j.m.stevens@uu.nl
[3] National Institute of Mental Health and Addiction, Utrecht, the Netherlands, {sdorsselaer,eduinhof,kmonshouwer}@trimbos.nl

Abstract. 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.

Keywords: Adolescent Alcohol Use · Exceptional Model Mining · Intersectionality · Local Pattern Mining · Validation of Subgroups · Trend Analysis

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 [18], and Monitoring the Future focuses on drug and alcohol use among students in America [10]. Using data of the Health Behaviour in School-Aged Children study (HBSC) [24] and the Dutch National School Survey on Substance Use (DNSSSU) [20], 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 [14] was followed by a sharp decrease between 2003 and 2015. Since 2015, both lifetime and monthly alcohol use among Dutch adolescents has remained stable [24,20]. The monthly prevalence of alcohol use in 2019 still ranges from 5% among 12-year-olds to 53% among 16-year-olds [20]. 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 [24].

Science is well aware that early and frequent alcohol use leads to a broad range of negative consequences [23,16]. 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. Hence, in this paper, we deploy Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) [21] as a sociological method.

EMM-RCS falls under the umbrella of Exceptional Model Mining (EMM) [6], 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 (in)sensitive to certain policy measures.

Although the current paper is closely related to [21], we believe to make new scientific contributions as follows: 1) the current paper discusses methodological challenges from a sociological point of view. Therefore, to ensure the validity of discovered subgroups, we discuss the complete result set before and after validation, 2) [21] discovered deviations in the course (i.e., change-over-time) of the trend in monthly alcohol use and exceptionally horizontal trends (i.e., no change-over-time). Here, we give more background information on how these subgroups can be discovered, and evaluate the findings from both a methodological and sociological point of view, 3) we consider an extra scenario where a single deviation of the prevalence of monthly alcohol use affects the entire trend line.

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 [8,19,15,14], 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 [4], 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 [3,9]. 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) [15], or with a structural equation model [8]. 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 [2]. For instance, the change-over-time of the monthly prevalence of alcohol use is assessed by repeatedly sampling new cases 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.

3 Data

Adolescent alcohol use in the Netherlands is monitored by two studies: the Dutch National School Survey on Substance Use (DNSSSU) [20] and the Health Behaviour in School-aged Children study (HBSC) [24]. 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).

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 paper 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 [22].

4 EMM-RCS deployment

Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) was introduced by [21] 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 by letting combinations of conditions on socio-demographic variables select adolescents from the full population. Such a combination is called a *description*, and could be as follows: *age = 12* $\wedge$ *life satisfaction 8-10* $\wedge$ *0 skipped classes* (where $\wedge$ should be read as *and*; all conditions should be met). 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 at the start of Section 5. For the purposes of the discussion here, it is merely important that many candidate subgroups are automatically generated.

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.

Denote data collected with a repeated cross-sectional research (RCS) design as $\Psi = (\Omega_{x_1}, \ldots, \Omega_{x_t}, \ldots, \Omega_{x_T})$, which 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 $r_{x_t} \in \Omega_{x_t}$ 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 [1], 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 $se(\mu_{x_t})$.To gauge whether a subgroup's trend in monthly alcohol use deviates from the population trend, EMM-RCS uses a generic quality measure [21]:

$$\varphi_{\text{RCS}}(D) = f\left(\{z_{x_t} \mid x_t \in \mathcal{T}\}\right) \tag{1}$$

$$z_{x_t} = \frac{|\theta^{SG}_{x_t} - \theta^{0}_{x_t}|}{se(\theta^{SG}_{x_t})}. \tag{2}$$

The quality measure consists of an inner part that measures exceptionality per year (Equation (2)) and an outer part that summarizes the T values into one overall quality value (Equation (1)).

4.1 Quantifying exceptionality

The following paragraphs detail how to instantiate Equations (1) and (2) to measure the three types of exceptionality that are relevant for better understanding adolescent alcohol use; that is, to discover subgroups of adolescents with 1) deviations in the prevalence of monthly alcohol use (new compared to [21]) 2) deviations in the course (i.e., change-over-time) of the trend in monthly alcohol use and 3) horizontal trends (i.e., no change-over-time) in monthly alcohol use.

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 $x_t \in \mathcal{T}$. In other words, as a statistic in Equation (2) we use the monthly prevalence of alcohol use. 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 that can be found 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; $se(\tau_{x_t}) = \sqrt{b^2_{x_t} se(\mu_{x_t})^2 + b^2_{x_{t-1}} se(\mu_{x_{t-1}})^2}$, 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 (as above). 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, this looks as follows: $f(\cdot) = f_{\text{countsum}}(\epsilon) = \sum \{\text{abs}(z_{x_t} - \epsilon) \mid x_t \in \mathcal{T}'', z_{x_t} < \epsilon\}$ with $|\mathcal{T}''| = T-2$ 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 5.3).

4.2 Validation measures

Although beam search is heuristically-guided by a quality measure, the exploratory nature of EMM-RCS may still require some additional measures to ensure that all discovered subgroups are valid and practically relevant. In this study, we propose to combine three validation techniques: dominance-based pruning, the distribution of false discoveries and a minimum improvement threshold.

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 [13]. This may happen especially with binary variables. For instance, the description $sex = female \wedge ethnic\ group = western$ 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 $age \leq 14 \wedge sex = female \wedge ethnic\ group = western$ may have been discovered. Here, we say that a subset of conditions (based on sex and ethnic group) *dominates* the original description (based on age, sex and ethnic group) [13].

We apply a form of dominance-based pruning (DP) 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. 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 DP 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 Second, the quality values of the top-q subgroups are compared against a *null distribution of false discoveries* (DFD) [7]. The DFD is constructed as follows. First, we randomly swap the values of socio-demographic variables between cases while keeping the information on alcohol use intact. For instance, 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, 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 socio-demographic variables on the one side and alcohol use on the other side.

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 swapped 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. It is then possible to run EMM-RCS 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 present study uses $m = 100$ and a one-sided significance level $\alpha_{DFD} = 0.025$. Note that if m is not sufficiently large, a p-value can be calculated non-parametrically by means of ranks.

Minimum improvement threshold per condition Third, we ensure that all conditions of a subgroup description substantially add to the quality of the subgroup by adding a minimum improvement threshold (MI). While DP removes conditions that decrease the quality value of a subgroup, here we evaluate whether every condition results in an increase in quality that is substantial enough to be considered meaningful in practice. We apply MI at the end of the search procedure. As a consequence, the removal of conditions may lead to the occurrence of similar subgroups and the final result set may become smaller than initially planned.

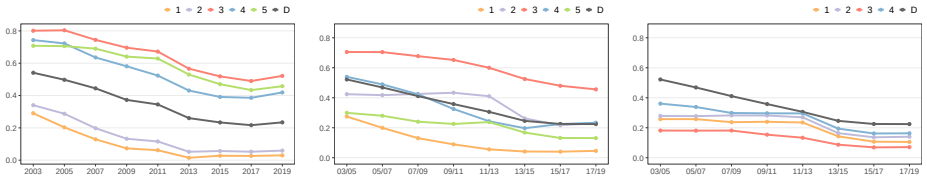
Of course, determining an appropriate threshold value is not straightforward. On the one hand, the threshold value depends on the scale of the quality measure. On the other hand, its interpretation is directly related to the domain of interest. In this context, we determined together with domain experts that proportions that increase with less than 1 percent point are not very relevant; the threshold value was set to $\delta = 0.025$ (2.5%).

5 Experimental Setup and Results

To traverse the candidate subgroup space, we employ *beam search* [6, Algorithm 1], with parameters $w = 20$, $d = 3$, and $q = 20$. One could set these parameters to higher values, to explore more of the candidate subgroup space. 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

Table 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 2).

(b) Exceptional slopes of *weighted moving average of the prevalence*: overall population and five exceptional subgroups (cf. Table 3).

(c) Exceptionally horizontal *weighted moving average of the prevalence*: overall population and four exceptional subgroups (cf. Table 4).

Fig. 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. The black line displays the population trend.

reduced interpretability of results. We dynamically discretize continuous variables using the `lbca` strategy with octiles from [17]. In addition to the three validation methods described above in Section 4.2, we apply two pruning strategies to increase the diversity of the result set: a weighted coverage scheme with $\gamma = 0.9$ [12,13] and description-based selection with a fixed size of $2w$ [13].

Table 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 is presented in black in Figure 1.

In the following three paragraphs, subgroups are manually split into trend groups (columns "TG") that share certain characteristics. From each such trend group, we take the top (most exceptional) subgroup, and we display its trend in the corresponding subfigure of Figure 1. Individual trends of the other subgroups in the tables are available on our interactive dashboard.[4]

5.1 Exceptional prevalence deviations

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 2. DP 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

[4] 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/`.

Table 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		
			condition 1	condition 2	condition 3
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	

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 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

Table 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		
			condition 1	condition 2	condition 3
1	1	0.19	age: 12		
	2	0.17	age: 12	ethn.: dutch, non-western	
	3	0.15	age: 12	complete family: yes	
	4	0.17	age: 12	ethn.: dutch, non-western	life satisf: 0-10
	7	0.16	age: 12	ethn.: 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	ethn.: dutch, non-western	
	15	0.54	age: 12-14	ethn.: 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 lvl: $\geq$ vmbo-p/t	
	14	0.32	age: 15-16		
	18	0.11	skipped classes: ≥ 1		
4	10	0.23	school lvl: $\geq$ havo	urbanity: at most moderate	
5	12	0.14	age: 15-16	urbanity: at least high	
6	16	0.13	school lvl: vmbo-p/t - havo	ethn.: (non-)western	

in time (Section 4.1), the prevalence in monthly alcohol use differs from the population prevalence at *all* waves (Figure 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.

5.2 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 3. DP 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 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 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 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 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 decrease 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 in subgroups 3 and 9 and the condition on skipped classes in subgroup 12 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.

5.3 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 before the application of any pruning strategies is listed in Table 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

Table 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		
			condition 1	condition 2	condition 3
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 lvl: $\leq$ havo/vwo
	6	0.08	ethnicity: non-western	age: 13-16	school lvl: $\geq$ vmbo-t
2	2	0.08	ethnicity: non-western	age: 14-16	urbanity: $\geq$ moderate
	15	0.09	ethnicity: non-western	age: 14-16	urbanity: $\geq$ 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 lvl: $\geq$ 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 lvl: $\leq$ havo/vwo
	10	0.14	age: 12	ethnicity: dutch	skipped classes: 0-6
	12	0.30	age: 12-13	life satisf: 7-10	school lvl: $\geq$ vmbo-t
	13	0.08	urbanity: very high	school lvl: $\leq$ havo	age: 13-16
	14	0.17	age: 12-13	sex: boy	school lvl: $\leq$ havo/vwo
	16	0.08	age: 12-13	life satisf: 9-10	school lvl: $\leq$ 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

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 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 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 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 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.

6 Discussion and 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. Beyond scientific interest, such insights have the potential to pinpoint directions for policies and interventions. We demonstrate the value of deploying Exceptional Model Mining for Repeated Cross-Sectional data (EMM-RCS) [21] as a sociological method.

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 5.1). Furthermore, EMM-RCS discovers interactions with 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 (Sections 5.2 and 5.3). A more in-depth understanding of subgroups of adolescents displaying stable alcohol use trends is important; such trends may be worrisome and warrant attention.

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 5.2). Here, the hypothesis that being a member of multiple social groups has a cumulative [3,9] or even an aggravated effect on adolescent alcohol use (a.k.a. the multiple jeopardy hypothesis) [5,11] cannot be accepted. In sum, 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.

Disclosure of Interests. The authors declare that they have no conflict of interest.

Acknowledgments. This work is part of the research program Data2People with project Exceptional and Deep Intelligent Coach (EDIC) and partly financed by the Dutch Research Council (NWO).

Author contributions statement R.S and E.D. initiated the idea of utilizing a state-of-the-art data mining approach to solve problems with traditional statistical methods in a social sciences context. G.S, S.D. and K.M. formulated concrete research questions regarding the existence of exceptional trends in adolescent alcohol use. R.S., M.P. and W.D. developed the model and computational framework. R.S. designed and executed the experiments and took the lead in writing the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Data availability and reproducibility To ensure data privacy, we cannot provide the raw data collected by the Health Behaviour in School-Aged Children study (HBSC) and the Dutch National School Survey on Substance Use (DNSSSU). HBSC is an international study carried out in collaboration with WHO/EURO. Two of our authors, Gonneke Stevens and Saskia van Dorsselaer, are principal investigators in the Netherlands. One of our authors, Karin Monshouwer, is the principal investigator of DNSSSU, the Dutch branch of the international ESPAD study. We provide a link to our dashboard and additional material such as descriptive information, missingness percentages per category per variable per year and a slice of the complete and incomplete dataset at `https://github.com/RianneSchouten/AlcoholTrends_HBSCDNSSSU_EMM/`. Naturally, we are willing to respond to any requests for more information about our research design, methodology and findings (please contact corresponding author).

References

1. Bethlehem, J.: Applied survey methods: A statistical perspective. John Wiley & Sons (2009)
2. Bryman, A.: Social research methods. Oxford University Press (2016)
3. Cole, E.R.: Intersectionality and research in psychology. American psychologist **64**(3), 170 (2009)
4. 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), 139–167 (1989)
5. Dowd, J.J., Bengtson, V.L.: Aging in minority populations: an examination of the double jeopardy hypothesis. Journal of Gerontology **33**(3), 427–436 (1978)
6. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (2016)
7. Duivesteijn, W., Knobbe, A.: Exploiting false discoveries: statistical validation of patterns and quality measures in subgroup discovery. In: Proc. ICDM. pp. 151–160 (2011)
8. Geels, L.M., Bartels, M., van Beijsterveldt, T.C., Willemsen, G., van der Aa, N., Boomsma, D.I., Vink, J.M.: Trends in adolescent alcohol use: effects of age, sex and cohort on prevalence and heritability. Addiction **107**(3), 518–527 (2012)
9. Ghavami, N., Katsiaficas, D., 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**, 31–73 (2016)

10. Johnston, L.D., Miech, R.A., O'Malley, P.M., Bachman, J.G., Schulenberg, J.E., Patrick, M.E.: Monitoring the future national survey results on drug use, 1975-2020: Overview, key findings on adolescent drug use. Institute for Social Research (2021)
11. King, D.K.: Multiple jeopardy, multiple consciousness: The context of a black feminist ideology. Signs: Journal of women in culture and society **14**(1), 42–72 (1988)
12. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. Journal of Machine Learning Research (MJLR) **5**, 153–188 (2004)
13. van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery (DAMI) **25**(2), 208–242 (2012)
14. de Looze, M.E., van Dorsselaer, S.A., Monshouwer, K., 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**, 90–101 (2017)
15. de Looze, M.E., Raaijmakers, Q., Bogt, T.t., 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), 69–72 (2015)
16. McCambridge, J., McAlaney, J., Rowe, R.: Adult consequences of late adolescent alcohol consumption: a systematic review of cohort studies. PLoS Med **8**(2), e1000413 (2011)
17. Meeng, M., Knobbe, A.: For real: a thorough look at numeric attributes in subgroup discovery. Data Mining and Knowledge Discovery (DAMI) **35**(1), 158–212 (2021)
18. 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)
19. Richter, M., Kuntsche, E., de Looze, M., Pförtner, T.K.: Trends in socioeconomic inequalities in adolescent alcohol use in Germany between 1994 and 2006. International journal of public health **58**(5), 777–784 (2013)
20. Rombouts, M., van Dorsselaer, S.A., Scheffers-van Schayck, T., Tuithof, M., Kleinjan, M., Monshouwer, K.: Jeugd en riskant gedrag 2019. Kerngegevens uit het Peilstationsonderzoek Scholieren. Trimbos-Instituut, Utrecht (2020)
21. Schouten, R.M., Duivesteijn, W., Pechenizkiy, M.: Exceptional model mining for repeated cross-sectional data (EMM-RCS). In: Proc. SDM. pp. 585–593 (2022)
22. Schouten, R.M., Duivesteijn, W., 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)
23. Sleet, D.A., Ballesteros, M.F., Borse, N.N.: A review of unintentional injuries in adolescents. Annual review of public health **31**, 195–212 (2010)
24. Stevens, G.W., van Dorsselaer, S.A., Boer, M., de Roos, S., Duinhof, E.L., ter Bogt, T.F., Van Den Eijnden, R.J., Kuyper, L., Visser, D., Vollebergh, W.A., et al.: HBSC 2017. Gezondheid en welzijn van jongeren in Nederland. Utrecht University (2018)