



Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining

19 NOV 2024

Presented by Rianne M. Schouten

Discovering exceptional trends in adolescent alcohol use in the Netherlands

Discovering exceptional trends in adolescent alcohol use in the Netherlands



Rianne M.
Schouten



Gonneke W.J.M.
Stevens



Saskia A.F.M.
van Dorsselaer



Elisa L.
Duinhof



Karin
Monshouwer



Mykola
Pechenizkiy



Wouter
Duivesteijn



Discovering **exceptional** trends in adolescent alcohol use in the Netherlands



Rianne M.
Schouten



Gonneke W.J.M.
Stevens



Saskia A.F.M.
van Dorsselaer



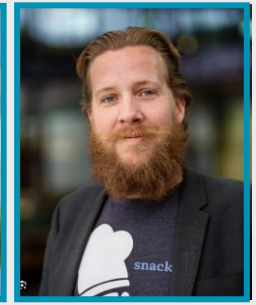
Elisa L.
Duinhof



Karin
Monshouwer



Mykola
Pechenizkiy



Wouter
Duivesteijn



Exceptional Model Mining for Hierarchical Data – a form of Local Pattern Mining

Discovering **exceptional** trends in adolescent alcohol use in the Netherlands



Rianne M.
Schouten



Gonneke W.J.M.
Stevens



Saskia A.F.M.
van Dorsselaer



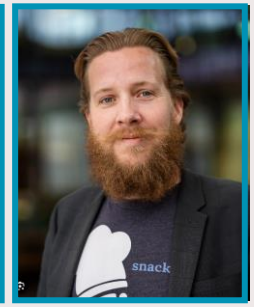
Elisa L.
Duinhof



Karin
Monshouwer



Mykola
Pechenizkiy



Wouter
Duivesteijn



Large study, repeated every 4 years,
in many countries in Europe

Discovering exceptional trends in adolescent alcohol use in the Netherlands



Rianne M. Schouten



Gonneke W.J.M. Stevens



Saskia A.F.M. van Dorsselaer



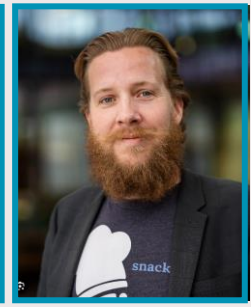
Elisa L. Duinhof



Karin Monshouwer



Mykola Pechenizkiy



Wouter Duivesteijn



Large study, repeated every 4 years, in many countries in Europe



Discovering **exceptional trends** in **adolescent alcohol use** in the Netherlands



Rianne M.
Schouten



Gonneke W.J.M.
Stevens



Saskia A.F.M.
van Dorsselaer



Elisa L.
Duinhof



Karin
Monshouwer



Mykola
Pechenizkiy



Wouter
Duivesteijn



Large study, repeated every 4 years,
in many countries in Europe

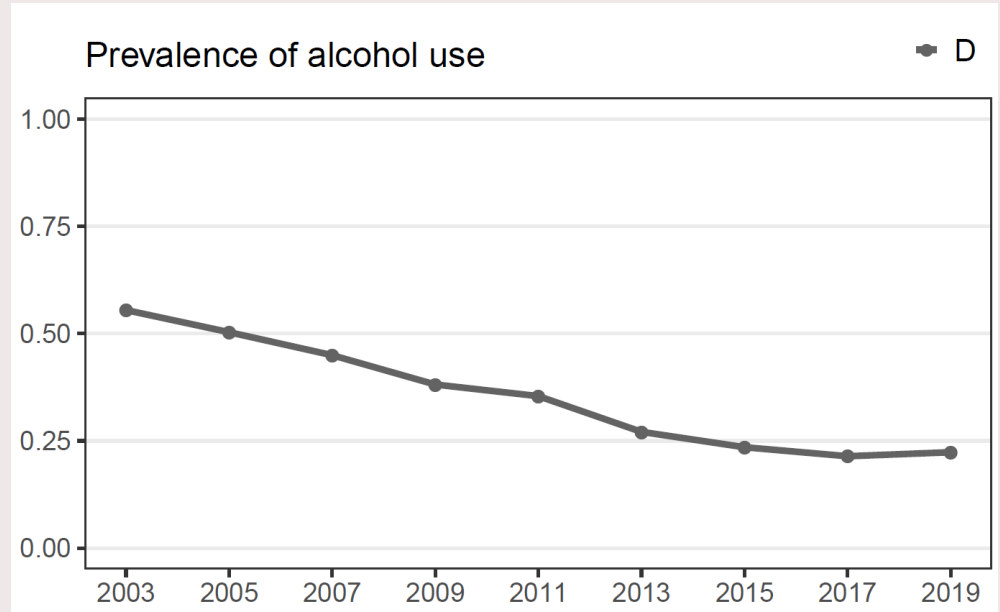


Population trend of adolescent alcohol use in the Netherlands

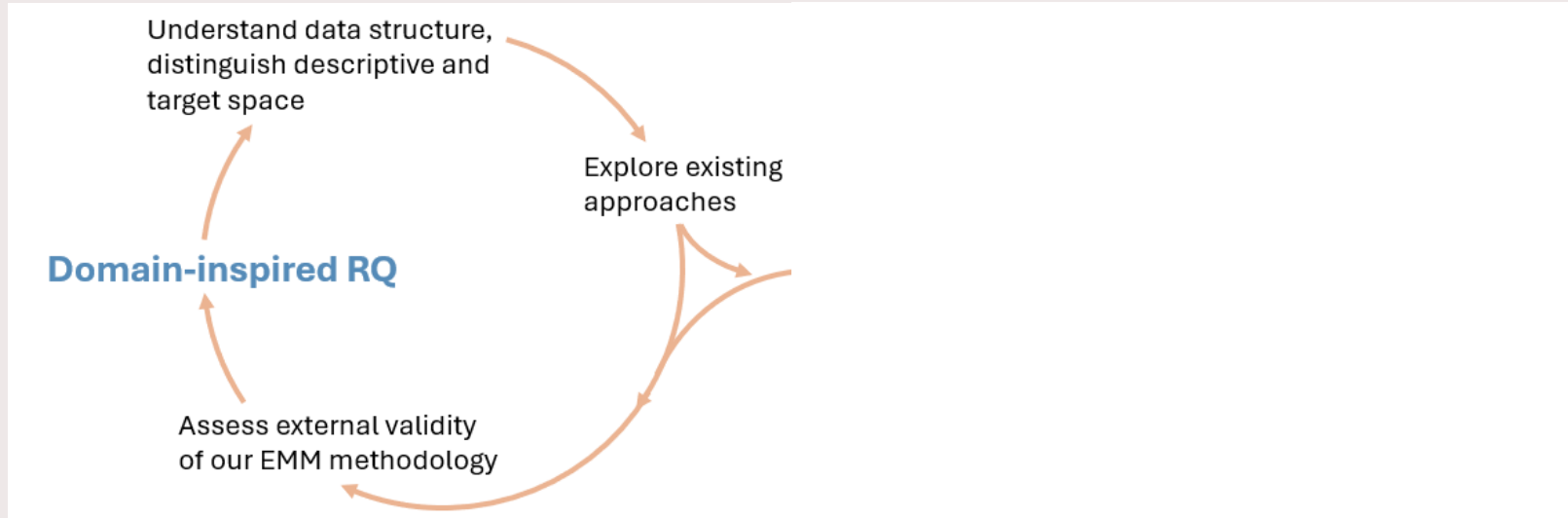
- 12–16-year-olds
- Prevalence of monthly alcohol use

Domain-specific RQ:

Are there subgroups of adolescents with exceptional trends in monthly alcohol use?

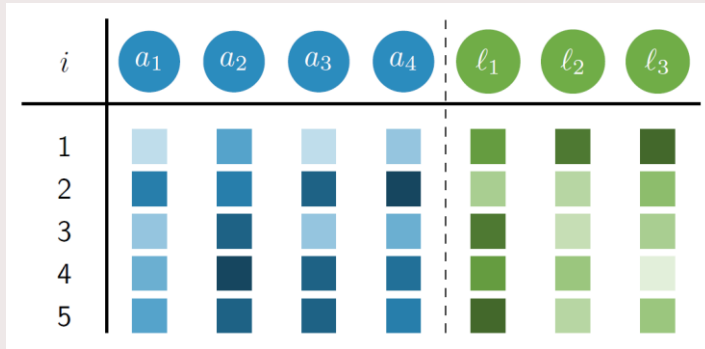


Research approach



Problems with existing approaches

In EMM: most methods developed for cross-sectional data.



A dataset Ω is a bag of n records $r^i \in \Omega$ of the form:

$$r = (a_1, a_2, \dots, a_k, l_1, l_2, \dots, l_m)$$

where k and m are positive integers.

We call a_1, \dots, a_k the **descriptors**, l_1, \dots, l_m the **targets**.

Description is conjunction of selection conditions on the descriptors, for instance:



What is our data structure?

Adolescent	School level	Year	Have you used alcohol in last 4 weeks?
1	VWO	2003	1
2	HAVO	2003	0
3	HAVO	2003	0
4	VMBO	2003	1
5	VWO	2005	1
6	VWO	2005	0
...			

This is called: **repeated cross-sectional data**

- Individuals may or may not recur between waves
- Sample sizes and distributions differ between waves
- Within a wave we assume independence, across waves it is unknown

Cross-sectional data in 2003

Cross-sectional data in 2005

Problems with existing approaches

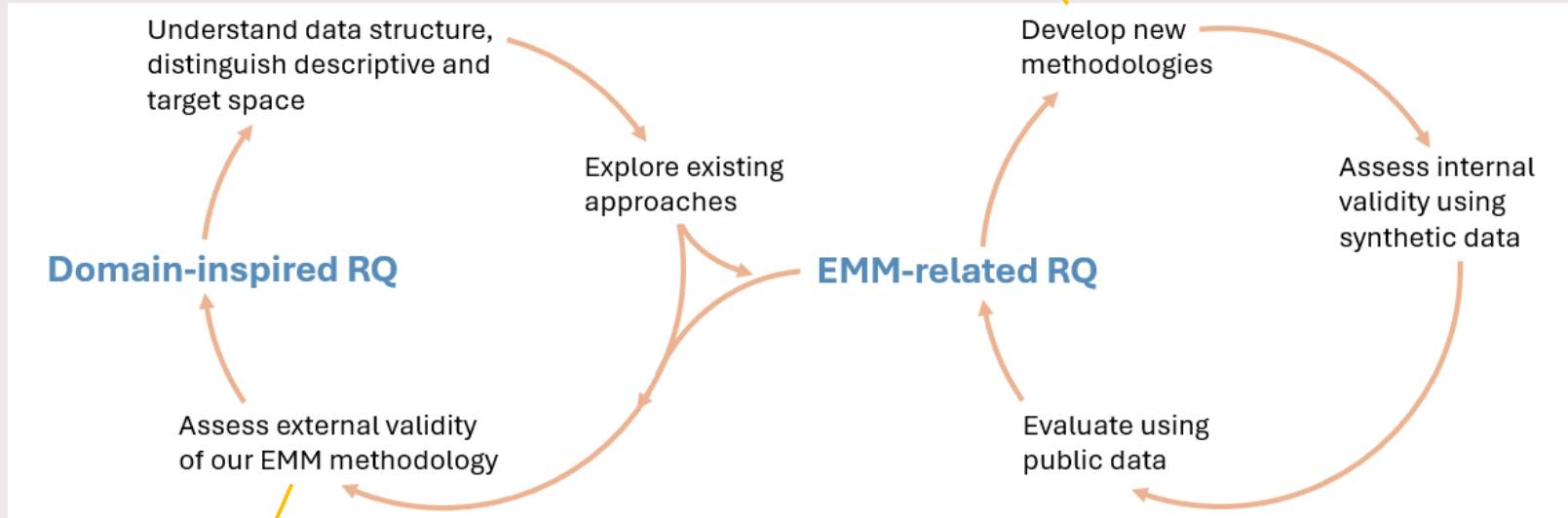
In EMM: most methods developed for cross-sectional data, but here we have repeated cross-sectional data with varying sample sizes and varying distributions over time.

With traditional statistical approaches:

1. Subgroups **hypothesized beforehand**, cannot include too many attributes, continuous attributes should be discretized
2. Particularly hard to test **combinations of conditions** (i.e., intersectionality theory), there are just too many
3. RCS design gives problems: trend represented with dummy attributes compares against a **reference year**, trend represented with a continuous attribute restricts the nature of relationship (i.e., **linear** in SEM)

Research approach

Schouten, R.M., Duivesteijn, W. & Pechenizkiy, M. (2022) [Exceptional Model Mining for Repeated Cross-Sectional Data \(EMM-RCS\)](#). In: Proc. SDM, pp. 585-593.



A generic quality measure to discover exceptional trends

Should say something about:

- Exceptionality of measurements **within** waves
- Aggregations of these exceptionalities **across** waves

$$\varphi_{RCS} = f(\{z_{x_t} \mid x_t \in T\}) \quad T = \{2003, 2005, \dots, 2019\}$$

Per wave, we calculate a z-score:
$$z_{x_t}(SG) = \frac{|\theta_{x_t}^{SG} - \theta_{x_t}^0|}{se(\theta_{x_t}^{SG})}$$

Sample size \uparrow = se \downarrow = $z_{x_t} \uparrow$ = $\varphi_{RCS} \uparrow$

Normalization by standard error achieves:

- Avoids tiny subgroups (an alternative for using the entropy function)
- Represents within-wave sample size, making scores comparable across waves
- Therefore helps with handling varying distributions over time

A generic quality measure to discover exceptional trends

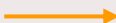
Should say something about:

- Exceptionality of measurements **within** waves
- Aggregations of these exceptionalities **across** waves


$$\varphi_{RCS} = f(\{z_{x_t} \mid x_t \in T\}) \quad T = \{2003, 2005, \dots, 2019\}$$

Aggregation can be: **sum, mean, max, count**, ..., depending on the type of trend deviation that is of interest


Three types of trend deviations

1. Exceptional deviations of the prevalence  new in this paper
2. Exceptional slope deviations
3. Exceptionally horizontal trends

Three validation measures

1. Dominance-based pruning (cf. van Leeuwen 2012)
 2. Distribution of False Discoveries (cf. Duivesteijn 2017)
 3. Minimum improvement threshold (cf. Bayardo 2000)
-  in this paper, we demonstrate the effects of these measures

Two other anti-redundancy strategies

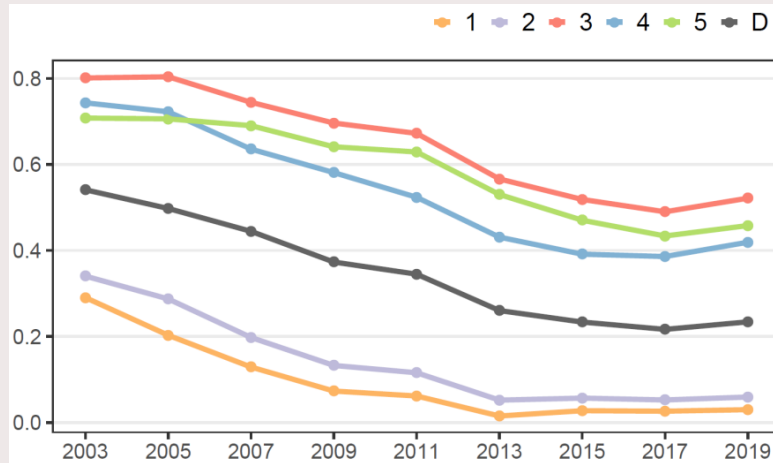
1. Fixed-size description-based selection (cf. van Leeuwen 2012)
 2. Weighted coverage scheme with $\gamma = 0.9$ (cf. van Leeuwen 2012)
-  The interaction between γ and beam-depth d is experimentally evaluated in:

[Schouten, R.M., Duivesteijn, W., Rasanen, P, Paul, J.M., & Pechenizkiy, M. \(2024\) Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression. In: Proc. ECML PKDD, pp. 66-82.](#)

Three types of trend deviations

1. Exceptional deviations of the prevalence (μ)

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005, \dots, 2019\}$$



Three types of trend devi

1. Exceptional deviations of the prevalence

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005\}$$

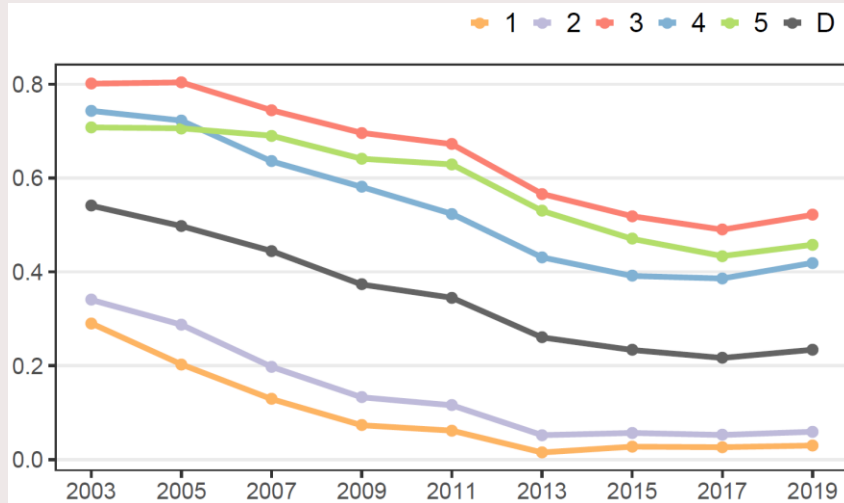


Table 2: Top-20 subgroups of adolescents with exceptional deviations of the prevalence of monthly alcohol use.

TG	SG	Cov	Description
			condition 1 condition 2 condition 3
1	0.11	age: 12	skipped classes: 0 urbanity: at least moderate
4	0.35	age: 12-13	skipped classes: 0 life satisf: 7-10
6	0.26	age: 15-16	ethnicity: dutch, western
11	0.48	age: 14-16	ethnicity: dutch, western
13	0.32	age: 15-16	

Three types of trend devi

1. Exceptional deviations of the prevalence

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005\}$$

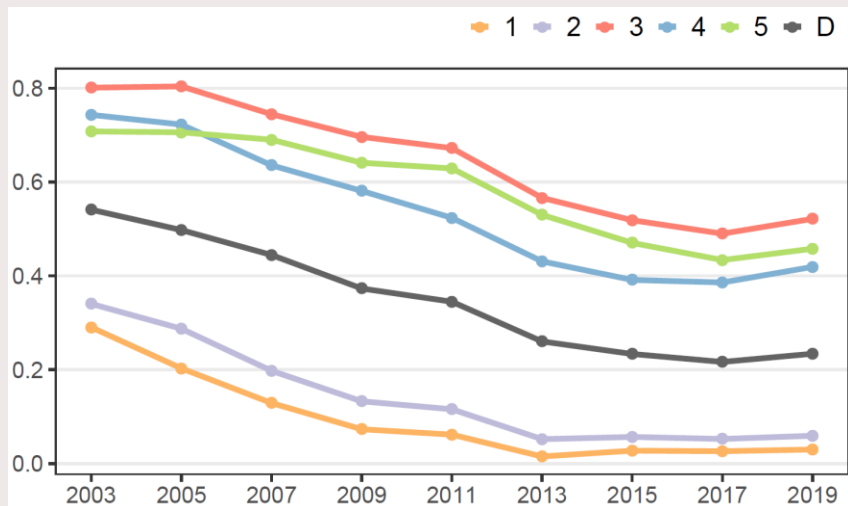


Table 2: Top-20 subgroups of adolescents with exceptional deviations of the prevalence of monthly alcohol use. Validation with a minimum improvement threshold results in the removal of 5 conditions (in red; the quality improvement is 1.4, 1.7, 0.4, -1.0 and 0.3 percent respectively). Three conditions 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
	4	0.35	age: 12-13	skipped classes: 0	life satisf: 7-10
2	6	0.26	age: 15-16	ethnicity: dutch, western	
	11	0.48	age: 14-16	ethnicity: dutch, western	
5	13	0.32	age: 15-16		

Three types of trend devi

1. Exceptional deviations of the prevalence

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005\}$$

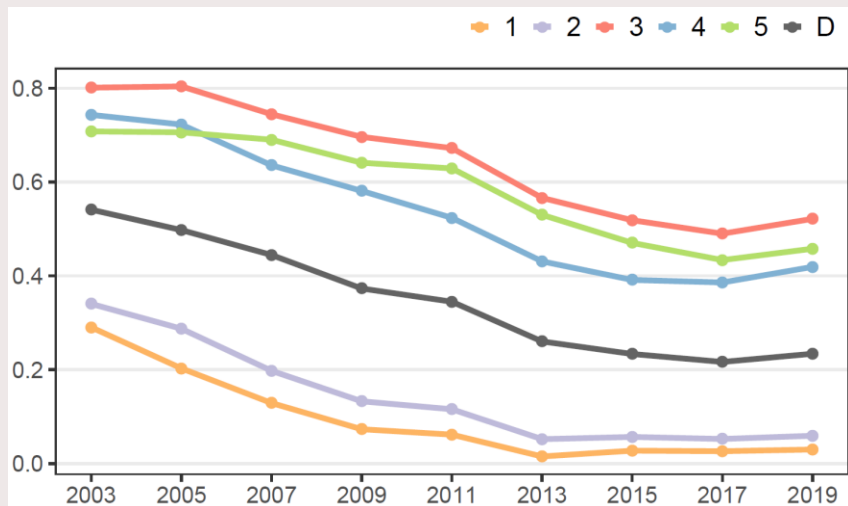


Table 2: Top-20 subgroups of adolescents with exceptional deviations of the prevalence of monthly alcohol use. Validation with a minimum improvement threshold results in the removal of 5 conditions (in red; the quality improvement is 1.4, 1.7, 0.4, -1.0 and 0.3 percent respectively). Three conditions 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	
	4	0.35	age: 12-13	skipped classes: 0	life satisf: 7-10	
2	5	0.37	age: 12-13	skipped classes: 0	life satisf: 6-10	
	7	0.40	age: 12-13	skipped classes: 0	urbanity: at least moderate	
	9	0.25	age: 12-13	life satisf: 6-10		
	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		

Three types of trend deviations

1. Exceptional deviations of the prevalence.

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005, \dots, 2019\}$$

2. Exceptional deviations of the prevalence.

$$\varphi_{RCS_2} = \max \frac{|\theta_{x_t}^{SG} - \theta_{x_t}^D|}{se(\theta_{x_t}^{SG})} : x_t \in \{2005, 2007, \dots, 2017\}$$

: $\tau_{x_t}^{SG}$ is the weighted moving average of two subsequent prevalences

: $\theta_{x_t}^{SG} = \tau_{x_{(t+1)}}^{SG} - \tau_{x_t}^{SG}$ θ is the slope of two subsequent averages

Three types of trend deviations

1. Exceptional deviations of the prevalence.

$$\varphi_{RCS_1} = \max \frac{|\mu_{x_t}^{SG} - \mu_{x_t}^D|}{se(\mu_{x_t}^{SG})} : x_t \in \{2003, 2005, \dots, 2019\}$$

2. Exceptional deviations of the prevalence.

$$\varphi_{RCS_2} = \max \frac{|\theta_{x_t}^{SG} - \theta_{x_t}^D|}{se(\theta_{x_t}^{SG})} : x_t \in \{2005, 2007, \dots, 2017\}$$

: $\tau_{x_t}^{SG}$ is the weighted moving average of two subsequent prevalences

: $\theta_{x_t}^{SG} = \tau_{x_{(t+1)}}^{SG} - \tau_{x_t}^{SG}$ θ is the slope of two subsequent averages

3. Exceptionally horizontal trends.

$$\varphi_{RCS_3} = \text{sum} \left\{ \text{abs}(z_{x_t} - \epsilon) \mid z_{x_t} < \epsilon \right\} \text{ with } z_{x_t} = \frac{|\theta_{x_t}^{SG} - \theta_{x_t}^D|}{se(\theta_{x_t}^{SG})} \text{ in this paper, } \epsilon \in \{0.005, 0.01, 0.02\}$$

Three types of trend deviations : validation

1. Exceptional deviations of the prevalence.

2. Exceptional slope deviations

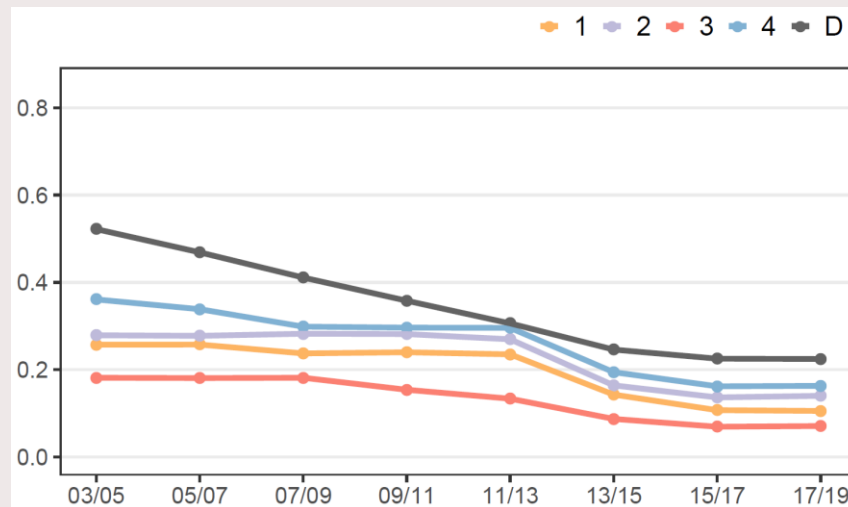
3. Exceptionally horizontal slopes

1. Dominance-based pruning (cf. van Leeuwen 2012)
2. Distribution of False Discoveries (cf. Duivesteijn 2017)
3. Minimum improvement threshold (cf. Bayardo 2000)

Exceptionally horizontal trends: DFD

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: \geq 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		



Main take-aways

- We developed a generic quality measure for discovering exceptional trends in **repeated cross-sectional data**: $\varphi_{RCS} = f(\{z_{x_t} \mid x_t \in T\})$
- 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.

Future work

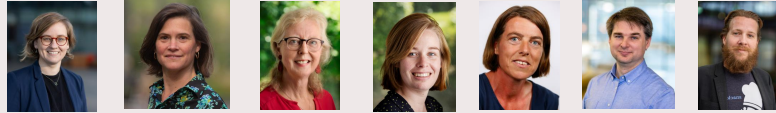
To further improve the real-world applicability of EMM, and LPM:

- Pruning and validation techniques work, but can we ground them in statistical theory?
- Focus on discovering subgroup sets rather than sets of subgroups?
- Applying EMM and LPM in a real-world scenario – adaptation to data streams?

Thank you!

r.m.schouten@tue.nl

<https://rianneschouten.github.io>



Schouten, R.M., Stevens, G.W.J.M., van Dorsselaer, S.A.F.M., Duinhof, E.L., Monshouwer, K., Pechenizkiy, M. & Duivesteyn, W. (2024) [Analyzing the interplay between societal trends and socio-demographic variables with local pattern mining: Discovering exceptional trends in adolescent alcohol use in the Netherlands](#). 2024.

Repository: https://github.com/RianneSchouten/AlcoholTrends_HBSCDNSSSU_EMM/

Interactive dashboard:

https://rianneschouten.shinyapps.io/InteractiveSupplement/?_ga=2.174229286.562920138.1629280969-190302859.1624274828/