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

Rianne Margaretha Schouten*  Wouter Duivesteijn*  Mykola Pechenizkiy*

Abstract
Repeted 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 people. To gain a deeper understanding of these trends, in [14] we proposed EMM-RCS, an Exceptional Model Mining instance designed to find subgroups displaying exceptional trend behavior in RCS data. We built 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. Here, we supplement the publication [14] with the following information. First, we explain our choice of search algorithm, parameter settings, anti-redundancy and validation techniques, and give a notion of the computational complexity of our method. Secondy, we provide extra synthetic data results. Thirdly, we discuss how threshold values can be evaluated using the Distribution of False Discoveries (DFD) [5]. Fourthly, we provide experimental results for an additional real-world dataset: the Eurobarometer. Finally, we provide further explanation and a demonstration of the working of a refinement strategy for incomplete descriptive attributes as introduced in [14, Definition 4.3].

1 Search algorithm
In our experiments, we choose to employ beam search [4, Algorithm 1] as algorithm to find exceptionally behaving subgroups. This is a heuristic search strategy. Exhaustive alternatives for EMM exist, but not without incurring a cost: optimality is guaranteed either by limiting the kinds of exceptional behavior that can be incorporated [1, 2, 8], or by static discretization of all numerical descriptors [10]. The versatility of our quality measure $\varphi_{RCS}$ [14, Section 4.2] disallows the first type of exhaustive search. Regarding the second type, in [14, Section 7], we find interesting subgroups defined in terms of numerical descriptors (e.g., age, life satisfaction). Beam search can determine suitable boundary values of the (half-)intervals restricting these numerical descriptors on the fly (which is possible with any method from [11] using local discretization timing; we employ the \texttt{lbca} strategy), which is to be preferred over a condition of the form age = high where high is predefined as an arbitrary threshold.

Beam search performs a level-wise search of $d$ levels: on each level, promising descriptions are further refined. Candidate subgroups are evaluated using $\varphi_{RCS}$: $w$ promising subgroups are taken to the next level for further refinement. The top-$q$ subgroups are returned.

EMM-RCS requires the computation of $T$ estimates and $T$ standard errors which is unlikely to be more complex than $O(Tn)$ for most statistical quantities ($n$ being the number of cases). The computational complexity of the entire beam search algorithm then becomes $O(dwZE(c + O(Tn) + \log(wq)))$, where we replaced the cost of learning a model $M$ from $N$ records on $f$ targets (cf. [4]) with $O(Tn)$, and where $Z$ and $E$ refer to the number of descriptors and the highest cardinality (number of distinct values) of any nominal attribute, respectively (numerical, ordinal and binary descriptors are refined faster than nominal descriptors).

Beam search requires a set of parameters whose settings may come across as arbitrary. In general, we choose the parameter values such that the result list is practical and meaningful. For the HBSC/DNSSSU dataset [14, Section 7.1], in consultation with domain experts we therefore choose $q = 20$ and $d = 3$. The latter will create descriptions with at most 3 conditions. For the public datasets (the Eurobarometer dataset in Section [4] and the Brexit dataset in [14, Section 7.2]), we use the same settings. Considering the number of descriptors and the total dataset size respectively, we set $w = 40$ and $csize = 0.05$ for the HBSC/DNSSSU dataset and the Eurobarometer dataset, and $w = 20$ and $csize = 0.1$ for the Brexit dataset. We only evaluate subgroups with a minimum proportion of available measurement occasions of $c_{occ} = 0.5$.

To reduce result set redundancy, we follow [7] in employing three techniques:

1. description-based selection with a fixed size of $2w =$
80 (subgroups are skipped if they have equal quality and the same description save for 1 condition);

2. fixed-size cover-based beam selection (a subgroup is weighted based on whether records are already covered by a subgroup with a higher quality [6]);

3. dominance pruning [7].

To validate our findings, we construct a Distribution of False Discoveries (DFD) [5] and compare the top-\( q \) subgroups against the distribution’s cutoff value for a one-sided significance level \( \alpha/2 = 0.025 \) [5]. To be precise, we create \( r = 100 \) random versions of a dataset by keeping descriptive space intact but taking a random permutation of rows in target space. We then run the beam search once for each of these datasets and construct a distribution of the quality values of the \( r \) top-1 subgroups. Because the swap randomization removes dependencies between descriptive and target space, the found subgroups are the consequence of chance and hence, high-quality false discoveries. We can then evaluate the quality value of a real subgroup against this distribution and reject the null hypothesis that the subgroup is a false discovery if it is larger than the cutoff value. All subgroups reported in this paper and in [14] survived this DFD validation: they are judged to be significantly different from false discoveries. Additionally, in Section 3 we will see that the comparison implicit in DFD validation also helps with setting threshold values.

### 2 Extra results for synthetic data experiment

Table 1 presents additional results for the synthetic data experiment of [14, Section 6]. Out of 100 repetitions per combination of simulation conditions, it presents the fraction of ground truth subgroups that is found in the top-20 result list, and the median rank and the median quality value of the found subgroups. The subgroup trend has a distance from the non-subgroup trend with \( \text{dist} \in \{1, 2, 3\} \), a standard deviation \( \text{sd} \in \{1, 2, 3\} \) and a ground truth description of \( \#\text{lits} \in \{2, 3, 4\} \) literals. The \#lits indirectly influences the subgroup size (25%, 12.5% and 6.25% respectively). Simulation is run with 100 repetitions.

<table>
<thead>
<tr>
<th>dist</th>
<th>sd</th>
<th>#lits</th>
<th>frac.found</th>
<th>med.rank</th>
<th>med.quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13.5</td>
<td>3</td>
</tr>
<tr>
<td>1 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>11.4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0.56</td>
<td>1</td>
<td>1</td>
<td>9.4</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>7.1</td>
<td>2</td>
</tr>
<tr>
<td>1 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6.4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.67</td>
<td>1</td>
<td>1</td>
<td>5.4</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0.07</td>
<td>10</td>
<td>10</td>
<td>6.0</td>
<td>4</td>
</tr>
<tr>
<td>3 3</td>
<td>0.92</td>
<td>5.5</td>
<td>5.5</td>
<td>4.8</td>
<td>3</td>
</tr>
<tr>
<td>4 4</td>
<td>0.47</td>
<td>3</td>
<td>3</td>
<td>4.2</td>
<td>5</td>
</tr>
<tr>
<td>1 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>26.0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0.52</td>
<td>1</td>
<td>1</td>
<td>17.6</td>
<td>1</td>
</tr>
<tr>
<td>2 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>13.4</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>0.64</td>
<td>1</td>
<td>1</td>
<td>9.4</td>
<td>2</td>
</tr>
<tr>
<td>3 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8.1</td>
<td>3</td>
</tr>
<tr>
<td>4 4</td>
<td>0.56</td>
<td>1</td>
<td>1</td>
<td>6.8</td>
<td>4</td>
</tr>
<tr>
<td>2 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>38.9</td>
<td>5</td>
</tr>
<tr>
<td>2 2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>33.0</td>
<td>6</td>
</tr>
<tr>
<td>4 4</td>
<td>0.66</td>
<td>1</td>
<td>1</td>
<td>26.8</td>
<td>7</td>
</tr>
<tr>
<td>2 2</td>
<td>1.0</td>
<td>2</td>
<td>2</td>
<td>19.8</td>
<td>8</td>
</tr>
<tr>
<td>4 4</td>
<td>0.58</td>
<td>1</td>
<td>1</td>
<td>17.0</td>
<td>9</td>
</tr>
<tr>
<td>3 3</td>
<td>0.99</td>
<td>4</td>
<td>4</td>
<td>13.3</td>
<td>10</td>
</tr>
<tr>
<td>4 4</td>
<td>0.94</td>
<td>1</td>
<td>1</td>
<td>11.7</td>
<td>11</td>
</tr>
<tr>
<td>4 4</td>
<td>0.63</td>
<td>1</td>
<td>1</td>
<td>9.4</td>
<td>12</td>
</tr>
</tbody>
</table>

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

\( \theta^\psi \) will therefore be smaller than the \( \text{dist} \) intended, depending on the size of the subgroup and the standard deviation in the subgroup. For instance, if the trend values of the subgroup and non-subgroup cases are both drawn from the same normal distribution and if \( \text{dist} = 1 \), for a subgroup with 25%, 12.5% or 6.25% coverage, the observed distance between \( \theta^{SG} \) and \( \theta^\psi \) will be 0.75, 0.875 and 0.9375 respectively. This effect increases if the standard deviation is larger in either the subgroup or non-subgroup. Therefore, under some simulation conditions, it can be challenging to find a larger subgroup and to give it a high rank.

### 3 Extra results for HBSC/DNSSSU dataset

In [14, Section 7.1], we searched for subgroups with exceptionally horizontal trends by setting \( \theta^{SG}_{x_i} \) as the slope between two moving averages [14, Equations (5.6) and (5.8)], by defining \( \theta^\psi = 0 \), by removing the correction with standard error \( (se(\theta^{SG}_{x_i}) = 1) \) and by counting the number of slopes within a certain threshold: \( \text{count}(\epsilon) = |x_i < \epsilon| \) [14, Section 5.2]. Originally we reported results for \( \epsilon = 0.01 \) in [14,
Due to the large scale of the survey, there are no descriptive attributes with no missing values. We drop all descriptive attributes with $\geq 50\%$ missing values, because it is likely that those attributes are surveyed in other years or that they result from follow-up questions (which are only asked depending on a respondent’s answer to another question). We apply \cite[Definition 4.3]{14} to the resulting 38 descriptors. Of those descriptors, 17 are binary, 1 numerical, 4 nominal, and 16 ordinal. In total, the dataset contains 155,244 records.

On the Eurobarometer dataset, we analyze the trend in the mean perception towards advancement of European unification. Figure 1 displays subgroups with exceptional trends obtained with summary functions $f_{\text{avg}}$ (left) and $f_{\text{max}}$ (right).

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

These findings are important from a domain perspective: they display a socio-demographic factor interplay that is difficult to find with confirmatory, global analysis techniques (cf. \cite[Section 2]{14}). From a data mining perspective, these findings illustrate that EMMRCS detects trends deviating from the global trend in both upwards and downwards directions.

Compared to Figure 1a, the subgroups in Figure 1b deviate from the general trend at some points (e.g., #18 in 1993 and #19 in 1992) but overlap with the general trend at other occasions. We find subgroups that barely exceed the minimum size constraint $c_{\text{size}} = 0.05$ (cf. Figure 1d). Found subgroups cover the citizens of specific countries. For instance, in Denmark (#19), citizens were more positive between 1992 and 1996 than the average European citizen while in France (#18), the perception of citizens dropped quickly between 1990 and 1992 and then gradually increased again.

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

The Eurobarometer dataset has missing values for many descriptive attributes and therefore, a subgroup may not have any observed values at some measurement occasions. We see this in Figure 1a by interrupted trend lines. Yet, because in EMM-RCS we define...
the constraint \( c_{occ} \) \cite{14, Section 4.1}, the subgroups are observed for at least 6 out of 12 occasions: the missing data does not prohibit detecting exceptional trends.

5 Incomplete Descriptive Attributes

In RCS data, descriptive attribute \( a_j \) could be missing for a specific record \( r_{x_i} \), but observed for another record \( r_{x_i'} \) at the same occasion \( x_t \) (\( i \neq i' \)). Popular missing data methods such as dropping incomplete cases and mean/median imputation are insufficient for solving this issue, for several reasons. The former may simply drop the entire dataset. In less extreme examples, both methods suffer under all forms of missingness: for missingness other than Missing Completely At Random (MCAR), they are known to give biased estimates; under MCAR, they give unrealistic standard errors \cite{3,9}, especially when the proportion of missing values is high (which is typical for RCS data). Moreover, it is hard to impute attribute \( a_j \) at occasion \( x_t \) if it is not sampled at all at that same occasion.

We therefore introduced an extra refinement operator in \cite{14, Definition 4.3}. If the missing data are MCAR, we expect the condition \( R_{a_j} = 0 \) to not appear in the descriptions of the top-\( q \) result list. After all, in case of MCAR every record has the same, fixed probability of being incomplete and consequently, the distribution of the missing values is similar to the distribution of the observed values. If the data are Missing At Random (MAR), information about missing values lies in the observed data (for other attributes). We expect those other attributes to appear in the descriptions. When the data are Missing Not At Random (MNAR), the information about the missing values is missing from the data. Here, we expect \cite{14, Definition 4.3} to be helpful. For more on MCAR, MAR, and MNAR, see \cite{3,12,16}.

It turns out that none of the subgroups found in the experiments selects missing values from a descriptive attribute as proposed in \cite{14, Definition 4.3}. To show that these findings are expected if the data is Missing Completely At Random (MCAR) or Missing At Random (MAR), we now present the results of an experiment where we artificially remove values from the variable \textit{hindsight} in the Brexit dataset (using techniques from \cite{15}). Recall that \textit{hindsight} is an important at-

<table>
<thead>
<tr>
<th>#</th>
<th>Cov.</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1 | 0.17 | benefit = \{benefit\} \land 
|   |      | epimp1 = \{fairly, very important\} \land 
|   |      | satismo = \{fairly, very satisfied\} |
| 3 | 0.11 | age \geq 23 \land 
|   |      | epimp = \{not at all, not very important\} \land 
|   |      | satismo = \{not very, fairly, very satisfied\} |
| 14 | 0.23 | benefit = \{not benefit\} |
| 20 | 0.06 | nation = \{Ireland\} \land married \neq \{separated\} |

(c) Coverage and description of the subgroups shown in Figure 1a

<table>
<thead>
<tr>
<th>#</th>
<th>Cov.</th>
<th>Description</th>
</tr>
</thead>
</table>
| 1 | 0.12 | epimp = \{not at all, not very important\} \land 
|   |      | married \neq \{refused\} \land 
|   |      | poldisc = \{occasionally, frequently\} |
| 4 | 0.06 | nation = \{Ireland\} |
| 10 | 0.07 | nation = \{Italy\} |
| 18 | 0.07 | nation = \{France\} |
| 19 | 0.07 | nation = \{Denmark\} |

(d) Coverage and description of the subgroups shown in Figure 1b

Figure 1: Subgroups displaying unusual perception towards advancement of European unification. Advancement is given with a value between 1 (standing still) and 7 (as fast as possible). The x-axis shows in which years the question is asked. We connect the values only if they are known for subsequent measurement occasions. Subgroups in the figure and table on the left are found with \( f_{avg} \), subgroups on the right are found with \( f_{max} \).
Table 2: Main condition in the description of the top-1 subgroup found on the Brexit dataset; \textit{hindsight} is artificially amputed from a given percentage of rows, for MCAR, MAR, and MNAR missingness mechanisms.

<table>
<thead>
<tr>
<th>%</th>
<th>MCAR</th>
<th>MAR</th>
<th>MNAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>h.sight = wr.</td>
<td>h.sight = wr.</td>
<td>h.sight = wr.</td>
</tr>
<tr>
<td>0.15</td>
<td>EURef = rem.</td>
<td>h.sight = wr.</td>
<td>EURef = rem.</td>
</tr>
<tr>
<td>0.25</td>
<td>EURef = rem.</td>
<td>EURef = rem.</td>
<td>EURef = rem.</td>
</tr>
<tr>
<td>0.35</td>
<td>EURef = rem.</td>
<td>EURef = rem.</td>
<td>h.sight = missing</td>
</tr>
<tr>
<td>0.45</td>
<td>EURef = rem.</td>
<td>EURef = rem.</td>
<td>h.sight = missing</td>
</tr>
</tbody>
</table>

For another 25% of the cases and an observed value for 50% of the cases in the Brexit dataset have \textit{hindsight} = \textit{wrong}, and that a missingness percentage of 25% therefore means that we have an observed value \textit{wrong} for 25% of the cases, a missing value for another 25% of the cases and an observed value \textit{right} for 50% of the cases.

When data is MCAR, cases are randomly amputed: both values \textit{wrong} and \textit{right} will be missing from \textit{hindsight}. With MCAR, the distribution of \textit{hindsight} will not change; only the number of observed values will reduce. We see in Table 2 that when 15% of the rows have a missing value, not enough observations remain to reliably estimate the trend of the proportion of people who want to leave the EU. Instead, attribute \textit{EURef16}, which has a correlation with \textit{hindsight} of about -0.2, will be used. With a small missingness percentage of 5%, the search is not affected by the missing values.

We generate MAR data by amputing \textit{hindsight} based on \textit{EURef16} values: cases with \textit{EURef16} = \textit{remain} are more likely to have missing \textit{hindsight} values than cases with \textit{EURef16} = \textit{leave}. Hence, the missing data is MAR since the information about the missing values is observed. Consequently, as expected, the top-1 subgroup is selected using \textit{EURef16} = \textit{remain} (cf. Table 2). Because the two attributes do not perfectly correlate, a MAR mechanism may ampute \textit{hindsight} = \textit{right} as well. It has been shown that MAR converges to MCAR for small correlations, and to MNAR for strong correlations [16]. Accordingly, for small missingness percentages, we find that \textit{hindsight} again appears in the subgroups’ descriptions.

With MNAR, the information about the missing values is really missing from the data. Therefore, for high missingness percentages, EMM-RCS needs to resort to \textit{[14, Definition 4.3]} to find the subgroup of non-leavers (cf. Table 2, right column, bottom two rows). For medium percentages, the subgroup of missing values is too small to compete with other subgroups: other attributes will be used for describing subgroups.

Acknowledgments

We gratefully acknowledge all people who put in a lot of effort to 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 Dorselaer for conducting the HBSC and DNSSSSU studies, for giving us access to a clean and well-constructed dataset, and 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, which is partly financed by the Dutch Research Council (NWO).

References


