Learning Prefix-Based Patterns from Demographic Sequences

Danil Gizdatullin[1], Jaume Baixeries[2], Dmitry I. Ignatov[1], Ekaterina Mitrofanova[1], Anna Muratova[1], Thomas H. Espy[3]

[1] National Research University Higher School of Economics, Moscow, Russia
{dignatov,emitrofanova,dgizdatullin}@hse.ru
[2] Universitat Politècnica de Catalunya, Barcelona
jbaixer@cs.upc.edu
[3] University of Pittsburgh, USA
the7@pitt.edu

Abstract. There are many different methods for computing relevant patterns in sequential data and interpreting the results. In this paper, we compute emerging patterns (EP) in demographic sequences using sequence-based pattern structures, along with different algorithmic solutions. The purpose of this method is to meet the following domain requirement: the obtained patterns must be (closed) frequent contiguous prefixes of the input sequences. This is required in order for demographers to fully understand and interpret the results.

Keywords: demographic sequences, pattern structures, sequence mining, emerging patterns, emerging sequences, machine learning

1 Introduction and related work

Demographic sequences are composed of vital facts that occur during the lifetime of a person, such as date of birth and first year of school. The analysis of these sequences is a popular and promising study direction in demography [1,2]. Among different results that can be obained by the analysis of demographic sequences, in this paper we focus on the computation of events that are relevant for characterising differences across generations. For example, would like to explore the main differences between consecutive generations are, in terms of demographic behaviour: number of children, age of marriage, incidence of divorce, etc.

Demographers and sociologists do not currently have a simple, unified methodology for the computation and interpretation of such event patterns, so different techniques are used: sequence analysis [3,4,5] and statistical methods [6,7,8,9,10]. Demographers have also started to show great interest in machine-learning and pattern-mining techniques [11] and other sophisticated sequence-analysis techniques [12]. Although many different methods have been developed, the methodology used in this field is not fully convergent with state-of-the-art sequence-mining techniques.

In the previous paper [13], we used the SPMF (Sequential Pattern Mining Framework) [14] for mining frequent sequences and finding relevant emerging patterns. However, this approach has a drawback: the results it yields are hard to interpret. Our work with demographers made us realize that it would be useful to find contiguous, prefix-based patterns, since they are interested in the full starting parts of people's life trajectories without gaps.

The main goal of this paper is to find emerging patterns that allow us to the discern demographic behaviour of different groups of people, with one important restriction, which is necessary to ensure the interpretability of our results: the obtained patterns must be (closed) frequent contiguous prefixes of the input sequences.

The paper is organised in a following way. We briefly determine the scope of this paper in Section 2. In Section 3, we describe our demographic dataset. Section 4 introduces basic definitions and prefix-based contiguous subsequences, in terms of pattern structures, combined with emerging patterns. Experimental results are reported in two subsections of Section 5. Finally, in Section 6 we provide the main concusions of this paper.

2 Problem Statement

In general terms, in this paper we compute a set of patterns that can characterise one group of subjects (a generation, a geographically defined group of people, a gender) with respect to other groups. As an example, we would like to be able to answer questions about the relevant differences between men and women or between generations in terms of demographic behaviour or questions about the emerging patterns that distinguish two consecutive generations.

We want to answer all these questions, inter alia, by mining emerging contiguous patterns.

However, it is important to note that we need to compute patterns that can be interpreted by field experts. This is the reason why classification methods like SVM and artificial neural networks must be discarded.

3 Materials

The dataset for the study is obtained from the Research and Educational Group for Fertility, Family Formation and Dissolution at the Higher School of Economics[4]. We use the three wave panel of the Russian part of the Generations and Gender Survey (GGS), which took place in 2004, 2007 and 2011[5]. The dataset contains records of 4,857 respondents (1,545 men and 3,312 women). The datset gender imbalance is caused by the panel nature of the data: survey

[4] `http://www.hse.ru/en/demo/family/`

[5] This part of GGS "Parents and Children, Men and Women in Family and in Society" is an all-Russian representative panel sample survey: `http://www.ggp-i.org/`

respondents' attrition is an uncontrollable process. That is why the representative waves combine to form a panel with a structure dissimilar to that of the general sample.

In the database, the following information is indicated for each person: date of birth, gender (male, female), generation, type of education (general, higher, professional), locality (city, town, village), religion (yes, no), frequency of religious event attendance (once a week, several times a week, minimum once a month, several times in a year or never). In addition, the database provides the dates of significant events in their lives, such as first job experience, completion of highest level education, leaving the parental home, first partnership, first marriage and birth of the first child. There are eleven generations: the first is of those born in 1930-1934, the last is of those born in 1980-1984.

4 Sequence mining and emerging patterns

4.1 Pattern structures in a demographic context

A *prefix-based contiguous subsequence* (or simply *prefix*) of a sequence $s = \langle s_1, \ldots, s_k \rangle$ of length $k' \leq k$ is the sequence $s_1 = \langle s'_1, \ldots, s'_{k'} \rangle$, where $s_i = s'_i$ for all $i \leq k'$. The *relative support*, $rsup_T(s)$, of a prefix s in a set of sequences T is the number of sequences in T that start with s divided by $|T|$.

For example, for sequence $\langle \{education\}, \{work\}, \{marriage\} \rangle$, the subsequence $\langle \{education\}, \{marriage\} \rangle$ is not a prefix-based contiguous subsequence. But $\langle \{education\} \rangle$, $\langle \{education\}, \{work\} \rangle$ and $\langle \{education\}, \{work\}, \{marriage\} \rangle$ are in fact prefix-based contiguous subsequences.

Pattern structures were introduced in [15] to analyse complex data with object descriptions given in non-object-attribute-value form, for example, chemical graphs, syntactic trees, vectors of numeric intervals and sequences. The usage of pattern structures for sequence mining has already been successfully demonstrated in [16].

Let $(S, (D, \sqcap), \delta)$ be a *pattern structure* related to our demographic problem, where S is a set of sequences, D is a set of possible descriptions of patterns with an associated intersection operator $\sqcap$ and operator $\delta(s)$ returns the description of sequence s from D. For example, if we have two sequences $s_1, s_2 \in S$, then

$$\delta(s_1) = \langle e_1^1, e_2^1, \ldots, e_n^1 \rangle \text{ and } \delta(s_2) = \langle e_1^2, e_2^2, \ldots, e_m^2 \rangle.$$

In our case, e_i^j is an event which happened in a person's lifetime.

Given two descriptions $d_1, d_2 \in D$, the intersection operator $\sqcap$ returns their maximal common prefix. To generate the maximal common prefix for a sequences subset, we use the Galois operator denoted by $\diamond$ that results in the itersection of the descriptions of the input sequences. For example, for s_1, s_2 such that $\delta(s_1) = \langle e_1^1, e_2^1, \ldots, e_n^1 \rangle$ and $\delta(s_2) = \langle e_1^2, e_2^2, \ldots, e_m^2 \rangle$,

$$\{s_1, s_2\}^\diamond = \delta(s_1) \sqcap \delta(s_2) = \langle e_1, e_2, \ldots, e_k \rangle$$

where k is maximal such that $e_i = e_i^1 = e_i^2$ for all $i \leq k \leq \min\{n, m\}$.

The operation $\diamond$ applied to description $d = \langle e_1, e_2, \ldots, e_k \rangle$ in our case is

$$d^\diamond = \{s \in S \mid d \sqsubseteq \delta(s)\},$$

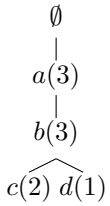
where $d \sqsubseteq \delta(s)$ means that d is a prefix of sequence s ($d \sqsubseteq \delta(s)$ if $d \sqcap \delta(s) = d$). In other words, the operator $\diamond$ applied to a description d returns the subset of objects in S that ha d as a prefix.

A pair (A, d) is a *pattern concept* of a pattern structure $(S, (D, \sqcap), \delta)$ with

$$A^\diamond = d \text{ and } d^\diamond = A, \text{ where } A \subseteq S, \text{ and } d \in D.$$

A is called the *extent* of the pattern concept (A, d) and d is its *intent*. For every pattern concept (A, d) of a pattern structure $(S, (D, \sqcap), \delta)$, it follows that $A^{\diamond\diamond} = A$ and $d^{\diamond\diamond} = d$. One may check that $(\cdot)^{\diamond\diamond}$ is a closure operator (idempotent, monotone, extensive) on 2^S w.r.t. $\subseteq$, so A is closed.

Let us discuss the representation of pattern concepts in the related prefix-tree (see section 4.5) for the original pattern structure and how they can be found. As an example, consider the set of sequences S, given by $\delta(s_1) = \langle a, b, c \rangle$, $\delta(s_2) = \langle a, b, c \rangle$ and $\delta(s_3) = \langle a, b, d \rangle$. The corresponding prefix tree is



We can extract the following pattern concepts relevant to this example: $(\{s_1, s_2, s_3\}, \langle a, b \rangle)$, $(\{s_1, s_2\}, \langle a, b, c \rangle)$, and $(\{s_3\}, \langle a, b, d \rangle)$. Any path of the associated prefix-tree, from its root to a bottom node, whose support is higher than the support of its descendants corresponds to the concept of an original pattern structure.

4.2 Hypotheses in pattern structures

Let us formulate a classification problem in a demographic setting. For each object (individual), there is also a target attribute (e.g. gender, for binary classification) according to which we want to classify that individual. Our pattern structure is then split into two pattern structures, positive $K_\oplus = (S_\oplus, (D, \sqcap), \delta)$ and negative $K_\ominus = (S_\ominus, (D, \sqcap), \delta)$, according to the target attribute which determines the class where it belongs to. Also, we have a set of undetermined sequences S_τ with unknown target attribute value.

Now, when the associated Galois operator is denoted as $A^\oplus$ for the positive pattern context and correspondingly for the negative one.

Let us define *positive* and *negative hypotheses*. A pattern intent of the pattern structure $K_\oplus$ ($K_\ominus$) $H \sqsubseteq D$ is a positive (negative) hypothesis if H is not a subset

of the pattern intent of any negative (positive) examples $s \in S_{\ominus}$ ($s \in S_{\oplus}$):

$$\forall s \in S_{\ominus}(s \in S_{\oplus}) : H \not\sqsubseteq s^{\ominus}(H \not\sqsubseteq s^{\oplus}).$$

Eventually, the hypothesis is the pattern intent of a pattern concept, which is found only in the objects of just one class.

4.3 Emerging patterns in pattern structures

Also, we introduce the notion of *emerging prefix-based contiguous subsequences* in terms of pattern structures. *Emerging pattern* is specific for one class, but not specific for its counterpart.

This feature is implemented via the ratio of the pattern supports for different classes. This ratio is called *growth rate*. The growth rate of a pattern $p \in D$ on positive and negative pattern structures of $K_{\oplus}$ and $K_{\ominus}$ is defined as

$$GR(p, K_{\oplus}, K_{\ominus}) = \frac{rsup_{K_{\oplus}}(p)}{rsup_{K_{\ominus}}(p)}$$

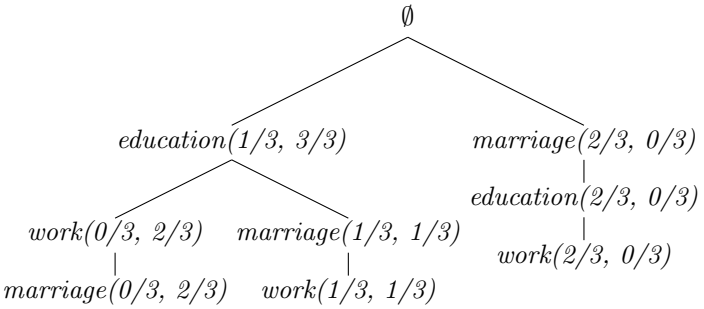
Patterns are selected by specifying a minimum growth rate as in [17]. That means, we set the minimum growth ratio, for which we want to select patterns:

$$GR(p, K_{\oplus}, K_{\ominus}) \geq \theta$$

Let us consider an example. Assume that we have two sets of sequnces.
Men's sequences:
$\langle\{education\}, \{work\}, \{marriage\}\rangle$
$\langle\{education\}, \{work\}, \{marriage\}\rangle$
$\langle\{education\}, \{marriage\}, \{work\}\rangle$

Women's sequences:
$\langle\{education\}, \{marriage\}, \{work\}\rangle$
$\langle\{marriage\}, \{education\}, \{work\}\rangle$
$\langle\{marriage\}, \{education\}, \{work\}\rangle$

So we can make a prefix-tree which based on these data.



On each node of the tree we store $rsup$ for different classes of current prefix-based contiguous subsequences. By using this structure we can compute growth rate.

4.4 Usage of emerging patterns for classification

For each class we compute its score as suggested in [18]. Let s be a new sequence which we want to classify, then its score in positive class is equal to

$$score_{\oplus}(s) = \frac{\sum_{p \in D^{\theta}_{\oplus}, p \sqsubseteq \delta(s)} GR(p, K_{\oplus}, K_{\ominus})}{median(GR(S_{\oplus}))}$$

where $D^{\theta}_{\oplus}$ is a subset of D that consists of all p with $GR(p, K_{\oplus}, K_{\ominus}) \geq \theta$. Then we choose all prefixes for the new sequence from set $D^{\theta}_{\oplus}$, and we sum all these growth-rates and get the score. Then we normalize the score by the median of the growth-rate for the current (positive) class.

4.5 Prefix-tree building: pseudocode and complexity

As we need to find contiguous prefix subsequences, it is a good idea to use a prefix-tree [19]. Usually, every node in a prefix-tree is associated with a certain string, but in our tree structure each node is associated with only one symbol (in case there are no simultaneous events).

Prefix-tree building. Let us start with the prefix-tree building procedure. As an input we have the set of sequences and their labels. At first we create the root node, which should be empty. Then we iterate through all the sequences and try to go the full path from the root to the end of a sequence. If we encounter a new path, we create a new node (and all the remaining events in the current sequence should be added as new subsequent nodes). Along the way, we increment all the counters associated with traversed nodes, one for each class.
Time complexity. Let n be the size of the training set and m be the number of different events in it. In line 4, we go through all the data; it takes n times. Then, in line 6, we go through all the elements in a sequence. The maximum length of a sequence is m. This pass takes $O(m)$ steps. Then in FIND procedure we iterate through all the children of a node. The maximum number of children nodes is $m - 1$; this step takes $m - 1$. Thus, the total time complexity is $O(n \cdot m^2)$. In our case, m is a small value (7-10 events) and can be considered a constant. The time complexity is $O(n)$.
Space complexity. The space complexity is equal to the number of nodes in a tree. The worst case is when all n sequences are different, i.e. all n sequences do not have the same prefix. In this case, space complexity will be $O(n \cdot m)$.

Algorithm 1 Sequence tree building

```
1: procedure SequencesTree(S, L)
2:     T ← {∅} // Initial prefix tree
3:     cn ← ∅
4:     for s ∈ S do
5:         l ← label(s)
6:         for e ∈ s do
7:             c ← Find(e, cn.children)
8:             if c ≠ ∅ then
9:                 c.counter[l] ← c.counter[l] + 1
10:                cn ← c
11:            else
12:                cn.children.append(newC)
13:                newC.element ← e
14:                newC.counter[l] ← 1
15:                cn ← newC

1: function Find(e, N)
2:     for n ∈ N do
3:         if n.element = e then return n
       return None
```

Algorithm 2 Classify Sequence

```
1: function ClassifySequence(T, s, l, Classes, minSup, minGR)
2:     sfc ← [0, 0]
3:     cn ← T.root
4:     for e ∈ s do
5:         for c ∈ cn.children do
6:             if c.element = e then
7:                 for l ∈ Classes do
8:                     if (c.support[l] > minSup) and (c.GR[label] > minGR) then
9:                         sfc[l] ← sfc[l] + n.GR[l]
10:                cn ← c
       return argMax(sfc)

1: procedure PrecomputeGrowthRate(T, Classes, soc)
2:     soc ←size(Classes) // soc is the number of classes
3:     for n ∈ T do // iterate over the tree nodes
4:         for l ∈ Classes do // iterate over the labels of classes
5:             n.support[l] ← n.counter[l]/soc[l]
6:     for n ∈ T do
7:         for l ∈ C do // GR is a growth-rate attribute
8:             n.GR[l] ← n.support[l]/n.support[counterpartL]
```

Classification by patterns. After performing the SequencesTree procedure on input data, we have the prefix-tree with absolute support value for each node and label. Then we can classify a new portion of sequences from the same domain. At first, we perform preprocessing via the PrecomputeGrowthRate procedure. In this procedure, we compute relative support and growth rates for each node. After that, we use the ClassifySequence function to predict the label of a new sequence.

Time complexity. Let k be the length of a sequence for classification. In PrecomputeGrowthRate, we need to iterate through the tree's nodes two times for each class label. We consider the situation with only two different classes: $O(n \cdot m \cdot 2) = O(n \cdot m)$.

In ClassifySequence, we iterate through the elements of the sequence and node children of nodes for each label: $O(k \cdot m \cdot 2) = O(k \cdot m)$.

5 Experiments and results

To perform experiments with pattern-based classification, we use Python and the Contiguous Sequences Analysis library implemented by the first author [6].

5.1 Classification by gender

After discussing with demographers, we have set the minimal relative support at 0.09. We have received the following prefix-based contiguous patterns that meet a minimum of 9% of all respondents 1, 2.

Table 1. Women's patterns

Pattern	Support
$\langle\{work\}\rangle$	0.287
$\langle\{work\}, \{education\}\rangle$	0.120
$\langle\{separation\}\rangle$	0.283
$\langle\{education\}\rangle$	0.239
$\langle\{education\}, \{work\}\rangle$	0.168
$\langle\{separation\}, \{education\}\rangle$	0.110
$\langle\{separation\}, \{education\}, \{work\}\rangle$	0.097

[6] `https://github.com/DanilGizdatullin/ContiguousSequencesAnalysis`

Table 2. Men's patterns

Pattern	Support
$\langle\{work\}\rangle$	0.329
$\langle\{work\},\{education\}\rangle$	0.155
$\langle\{separation\}\rangle$	0.266
$\langle\{education\}\rangle$	0.276
$\langle\{education\},\{work\}\rangle$	0.103
$\langle\{separation\},\{education\}\rangle$	0.199
$\langle\{separation\},\{education\},\{work\}\rangle$	0.099

After thoughtful inspection, we can conclude that the beginning of human life trajectories do not depend strongly on gender; moreover, the beginnings of the most popular paths are the same for both sexes. We have split all our data into two groups: a training set and a test set with the percentage of 66.5%-33.5%.

We have selected the same minimum support threshold for both classes, 0.004; this means that the pattern must be found in life trajectories of at least five men and nine women. Then we made a classification with different minimum threshold values for growth rates $\{1.5, 2, 5, 7\}$ for men and $\{1.5, 2, 5, 7, \infty\}$ for women.

The graphs below show the results and skyline in TPR-FPR (true positive rate, false positive rate), TPR-NCPR (true positive rate, non-classified positive rate), NCPR-FPR (non-classified positive rate, false positive rate) on the axes (Fig. 1 and 2).

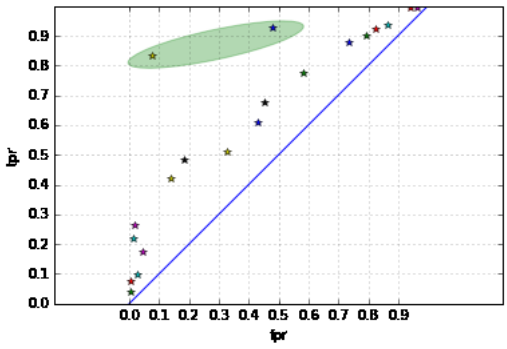


Fig. 1. TPR-FPR plot, along with two Pareto-optimal results from the skyline in the oval.

We have chosen the same value for both minimum support classes at 0.004. This means that the pattern must occur for at least five men and nine women. We have performed several classifications with different minimum values of the

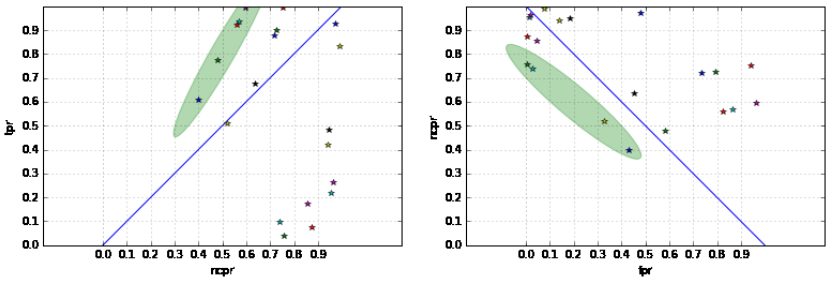


Fig. 2. TPR-NCPR (left) and NCPR-FPR (right) plots, along with their skylines in the ovals.

growth rate from $\{1.2, 1.5, 2, 3, 5, 7, \infty\}$. In a demographic setting, it is important to identify interesting discriminative patterns, though we do not try to solve the problem of classification by gender. Thus, many objects from the test set have not been assigned to any class. For example, in the experiment with the best TPR-FPR metric, we cover over 1% of people in the test sample. We can conclude from the obtained results that the interesting discriminative patterns for some class relative to another have a small cover. Moreover, we can conclude that the average behaviour of men and women has no stark differences in general, but there are local groups of both classes that behave sufficiently differently.

The best quality of classification has been reached with the minimum value of the growth rate (7, ∞). It corresponds to the following emerging patterns 3,4.

Table 3. Women's patterns in the test set

Pattern	Growth rate	Support
$\langle\{work, separation\}, \{marriage\}, \{children\}, \{education\}\rangle$	∞	0.006
$\langle\{separation, partner\}, \{marriage\}\rangle$	∞	0.005
$\langle\{separation, partner\}, \{marriage\}\rangle$	∞	0.005
$\langle\{work, separation\}, \{marriage\}, \{children\}\rangle$	∞	0.008
$\langle\{work, separation\}, \{marriage\}\rangle$	∞	0.009

Table 4. Men's patterns in the test set

Pattern	Growth rate	rsup
$\langle\{education\},\{marriage\},\{work\},\{children\},\{separation\}\rangle$	10.6	0.006
$\langle\{education\},\{marriage\},\{work\},\{children\}\rangle$	12.7	0.007
$\langle\{education\},\{work\},\{partner\},\{marriage\},\{separation\},\{children\}\rangle$	10.6	0.006

The presented results for women show that they prefer to start their adult lives with separation. Only in one case did separation coincide with having a partner. In other cases, we have an image of an independent woman, who got a job and left her parents. The second step in all the cases is marriage. Here we see how an independent (from parents and financially) woman creates her own family and bears a child. The longest sequence contains the event "finishing an education of the highest level". Only after fulfilling four important socioeconomic and sociodemographic events does the typical woman finish her education.

The first event in the sequences indicative for men is education. Unlike women, who obtain their education only after fulfilling almost all events, men are getting their education earlier. This can show not only the priority of education for men and women, but also the difference in the level of the highest step of the finished education: the lower the level, the lower the age of obtaining education.

The second event in the three sequences specifying men is marriage (two cases) or work (one case). Like a woman, a man tends to create his family quite early; unlike a woman, who is already very independent at this step, a man has only his education. A man whose second event is "marriage" obtains his first job next and then becomes a father. In the longest sequence, a man leaves the parental home as the final step in the transition to adulthood.

Men who have "work" as the second event demonstrate different events in their sequences: they have the first partner, then they get married, leave the parental home and – only after all other events – become parents.

5.2 Classification by generation

In this experiment we search for emerging patterns for different generations of the same sex. The first class 0 features people who were born between 1924 and 1959. The second class 1 contains people who were born between 1960 and 1984.

First, let us find emerging patterns for women from different generations. We have 940 women from class 0 and 1,364 women from class 1. We need to tune two parameters: the first is the minimal support and the second is minimal growth rate.

Let us tune the minimal support parameter (Table 5).

Table 5. Tuning of minimal support for women

minsup	accuracy	TPR	FPR	NCR non-classification rate
0.001	0.682	0.707	0.331	0.255
0.004	0.683	0.703	0.316	0.333
0.01	0.668	0.710	0.332	0.399
0.025	0.660	0.648	0.298	0.540
0.04	0.660	0.616	0.278	0.606
0.05	0.652	0.646	0.312	0.641
0.1	0.651	1.0	1.0	0.884

As we can see, the minimal support can sufficiently change only the non-classification rate and slightly affect accuracy, the TPR and the FPR.

We have chosen 0.004 as the minimal support and tuned minimal growth rate.

Table 6. Tuning of minimal growth rate for women

minGrowthRate	accuracy	TPR	FPR	NCR
1.5	0.683	0.655	0.297	0.102
2	0.692	0.703	0.316	0.333
3	0.766	0.747	0.217	0.684
5	0.751	0.821	0.347	0.848
7	0.777	0.848	0.333	0.891

We have decided to choose a minGrowthRate of 2, since it covers 0.66 percent of the test data and provides good results in terms of accuracy, the TPR and the FPR.

Since we have had many emerging patterns in the data, we consider only patterns with the greatest growth rate and support.

Table 7. Patterns for women of older generations

Pattern	Growth rate	Support
$\langle\{work\}, \{separation\}\rangle$	1.85	0.38
$\langle\{work\}, \{marriage, separation\}\rangle$	3.92	0.08

Table 8. Patterns for women of younger generations

Pattern	Growth rate	Support
$\langle\{education\}\rangle$	1.84	0.26
$\langle\{education\},\{work\}\rangle$	3.92	0.08

As we can see from Tables 7-8, the main differences in the behaviour of women from different generations lie in the tendency to obtain education and the tendency to work, and only after that separate from their parents in older generations.

Let us find emerging patterns for men from different generations. Again we should tune the minimal support:

Table 9. Tuning of minimal support for men

minsup	accuracy	TPR	FPR	NCR
0.001	0.701	0.667	0.266	0.271
0.004	0.704	0.667	0.262	0.338
0.01	0.723	0.671	0.232	0.442
0.025	0.719	0.651	0.218	0.590
0.04	0.706	0.536	0.165	0.712
0.05	0.718	0.627	0.208	0.764
0.08	0.710	0.0	0.0	0.944

Again, minimal support can sufficiently change only the non-classification rate.

We have chosen 0.01 as the minimal support and tuned minimal growth rate.

Table 10. Tuning of the minimal growth rate for men

minGrowthRate	accuracy	TPR	FPR	NCR
1.2	0.638	0.510	0.242	0.050
1.5	0.670	0.591	0.260	0.171
2	0.723	0.671	0.232	0.442
3	0.754	0.627	0.144	0.664
5	0.744	0.625	0.152	0.845
7	0.836	0.808	0.138	0.901

The patterns with the biggest growth rate and support are reported in Tables 11,12.

Table 11. Patterns for men of older generations

Pattern	Growth rate	Support
$\langle\{work\}, \{marriage, separation\}, \{education\}\rangle$	13.52	0.025
$\langle\{work\}, \{marriage\}, \{separation\}\rangle$	22.87	0.042
$\langle\{work\}, \{marriage\}, \{separation\}, \{education\}\rangle$	∞	0.0208

Table 12. Patterns for men of younger generations

Pattern	Growth rate	Support
$\langle\{education\}, \{work\}, \{separation\}, \{marriage\}, \{children\}\rangle$	10.58	0.020
$\langle\{education\}, \{work\}, \{separation, partner\}, \{marriage\}\rangle$	8.65	0.016
$\langle\{education\}, \{marriage, separation\}\rangle$	7.69	0.015

As in the previous experiment with the women subsample, the main difference lies in the tendency to obtain education; thus, men of younger generation demonstrates this tendency.

6 Conclusion

The main result of our work is the application of various pattern-mining approaches, including pattern structures, to the analysis of demographic sequences. The following conclusions can be drawn from the first results of this work:

1. In this paper, the application of sequence-based patterns for problems of demographic trajectories has been studied.
2. A new method based on pattern structures for the analysis of special pattern type required by demographers (prefix-based and contiguous) has been proposed and implemented.
3. Behavior patterns for different classes of respondents were obtained and interpreted for the most recent and clean demographic material available for Russia.
4. Classifications based on pattern structures and emerging patterns have been designed and tested.

According to the demographers involved in the project, the work is very important for the further development of the pattern-mining application for demographic analysis of sequence data. Thus, among the next planned steps are the following:

– using similarity [20] and kernel-based approaches [21] for demographic-sequence mining;

- (sub)sequence clustering, in particular, based on pattern structures;
- pattern-mining and rule-based approaches for next-event prediction [13] competitive with black-box approaches like recurrent neural networks;
- comprehensive trajectory visualisation within cohorts [22];
- analysing sequences of statuses like $\langle\{studying, single\}, \{working, single\}\rangle$;
- analysis of matrimonial and reproductive biographies, migration studies, etc.

Acknowledgments We would like to thank our colleagues Sergei Kuznetsov, Alexey Buzmakov, and Mehdi Kaytoue for their advice on pattern structures and sequence-mining, as well as Guozhu Dong for the interest in our work with emerging patterns and our colleagues from Institute of Demography at the National Research University Higher School of Economics.

This article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2016-2017 (grant 16-05-0011 Development and testing of demographic sequence analysis and mining techniques) and by the Russian Academic Excellence Project "5-100".

References

1. Aisenbrey, S., Fasang, A.E.: New life for old ideas: The second wave of sequence analysis bringing the course back into the life course. Sociological Methods & Research **38**(3) (2010) 420–462
2. Billari, F.C.: Sequence analysis in demographic research. Canadian Studies in Population **28**(2) (2001) 439–458.
3. Aassve, A., Billari, F.C., Piccarreta, R.: Strings of adulthood: A sequence analysis of young british womens work-family trajectories. European Journal of Population **23**(3/4) (2007) 369–388
4. Braboy Jackson, P., Berkowitz, A.: The structure of the life course: Gender and racioethnic variation in the occurrence and sequencing of role transitions. Advances in Life Course Research (9) (2005) 55–90
5. Worts, D., Sacker, A., McMunn, A., McDonough, P.: Individualization, opportunity and jeopardy in american womens work and family lives: A multi-state sequence analysis. Advances in Life Course Research **18**(4) (2013) 296–318
6. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociological Methods & Research (2000)
7. Billari, F., Piccarreta, R.: Analyzing demographic life courses through sequence analysis. Mathematical Population Studies **12**(2) (2005) 81–106
8. Billari, F.C., Fürnkranz, J., Prskawetz, A.: Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. European Journal of Population **22**(1) (2006) 37–65
9. Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C.: How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data. Sociological Methods & Research **38**(1) (2009) 197–231
10. Ritschard, G., Oris, M.: Life course data in demography and social sciences: Statistical and data-mining approaches. Adv. in Life Course Research **10** (2005) 283–314

11. Blockeel, H., Fürnkranz, J., Prskawetz, A., Billari, F.C.: Detecting temporal change in event sequences: An application to demographic data. In: Principles of Data Mining and Knowledge Discovery, 5th Eur. Conf., PKDD 2001. (2001) 29–41
12. Gabadinho, A., Ritschard, G., Mller, N.S., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR. J. of Stat. Software **40**(4) (4 2011) 1–37
13. Ignatov, D.I., Mitrofanova, E., Muratova, A., Gizdatullin, D.: Pattern mining and machine learning for demographic sequences. In: Knowledge Engineering and Semantic Web, KESW 2015, Proceedings. (2015) 225–239
14. Fournier-Viger, P., Lin, J.C., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H.T.: The SPMF open-source data mining library version 2. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III. (2016) 36–40
15. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: Harry S. Delugach and Gerd Stumme, editors, Conceptual Structures: Broadening the Base, volume 2120 of Lecture Notes in Computer Science. (2001)
16. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, A., Raïssi, C.: On Projections of Sequential Pattern Structures (with an Application on Care Trajectories). In: 10th Int. Conf. on Concept Lattices and Their Applications. (2013) 199–208
17. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: classification by aggregating emerging patterns. In: Discovery Science, 2nd Int. Conf., DS '99. (1999) 30–42
18. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proc. of the Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. KDD '99, ACM (1999) 43–52
19. de la Briandais: File searching using variable length keys. In: File searching using variable length keys. Proc. Western J. Computer Conf. pp. 295-298. Cited by Brass. Rene. (1959)
20. Egho, E., Raïssi, C., Calders, T., Jay, N., Napoli, A.: On measuring similarity for sequences of itemsets. Data Min. Knowl. Discov. **29**(3) (2015) 732–764
21. Elzinga, C.H., Wang, H.: Versatile string kernels. Theoretical Computer Science **495** (2013) 50 – 65
22. Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesøe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J., Brunak, S.: Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature Communications **5** (06 2014) 4022 EP –