Frequent sequence mining on longitudinal data

Isak Hietala
Master Thesis in Statistics and Data Mining

Frequent sequence mining on longitudinal data

Segregation of Swedish employees

Isak Hietala

Division of Statistics
Department of Computer and Information Science
Linköping University
Supervisor
Linda Wänström

Examiner
Oleg Sysoev
I shall try not to use statistics as a drunken man uses lamp-posts, for support rather than for illumination
- Andrew Lang
## Contents

Abstract .............................................. 1

Acknowledgments ...................................... 3

1. Introduction ....................................... 5
   1.1. Institute for Analytical Sociology .................. 5
   1.2. Background ......................................... 5
   1.3. Objective .......................................... 6
   1.4. Related work ....................................... 7

2. Data ................................................ 9
   2.1. Data sources ....................................... 9
       2.1.1. Databases .................................... 9
   2.2. Data management .................................... 11
       2.2.1. Discretization .................................. 13
       2.2.2. Sub-setting the data ............................ 13

3. Methods ............................................. 15
   3.1. Frequent sequence mining (FSM) ..................... 15
       3.1.1. Apriori property ............................... 18
       3.1.2. Method and algorithms ........................ 18
       3.1.3. Pruning frequent sequences .................... 21

4. Results ............................................. 23
   4.1. Data visualization ................................... 23
   4.2. Sequence mining .................................... 25
       4.2.1. Multiple events in an element .................. 25
       4.2.2. Single event in an element ...................... 26
       4.2.3. General results ................................ 27

5. Discussion ......................................... 33
   5.1. Evaluation of the results .......................... 33
   5.2. Evaluation of the methodology ...................... 34
   5.3. Further study ...................................... 35

6. Conclusions ........................................ 37
Abstract

This thesis is based on longitudinal data of the Swedish population provided by Statistics Sweden and is conducted on behalf of the Institute for Analytical Sociology. The focus is on investigating the effectiveness of a frequent sequence mining method called Constrained Sequential PAttern Discovery using Equivalence classes (cSPADE). The method is applied to data on segregation within workplaces, specifically reasons for Swedish employees moving to more segregated workplaces. The thesis found that no unique pattern of age, gender, education, unemployment, income, workplace size or foreignness index explain why a Swedish employee moves to a more segregated workplace. Evaluating the algorithm, it was found that the number of observations need to be smaller or an alteration of the algorithm needs to be done to reduce the process time for this specific data set.

Keywords: Longitudinal data, frequent sequence mining, cSPADE, segregation
Acknowledgments

Firstly I would like to thank my supervisor Linda Wänström at Linköping University for the valuable discussions, revisions and opinions given to me throughout the process. I would also like to thank Selcan Mutgan, Frederik Witte and Peter Hedström from the Institute for Analytical Sociology for the help in understanding the data set and any topical problems I have encountered. I also thank my opponent Sowmya Krishnaraj for reviewing the thesis and coming with helpful comments. Credit should also be given to Lennart Hietala, Lotta Järvstråt, Josefina Lundmark, and Henrik Olofsson for taking the time to review the thesis, as well as discuss ideas and problems encountered throughout the project.
1. Introduction

This chapter describes the background, objective of the thesis and some work related to segregation and longitudinal data.

1.1. Institute for Analytical Sociology

The Institute for Analytical Sociology (IAS) is a recently created institute located in Norrköping and part of Linköping University. The institute focuses on research in social, political and cultural areas with its researchers originating from many different fields.

1.2. Background

Segregation is, according to the Merriam Webster dictionary, the isolation of a group with specific characteristics from a larger group. The most prominent example of segregation over the last century is racial segregation, e.g. apartheid. Two different types of segregation exist, de jure (by law) and de facto (by practice). De jure segregation is controlled by laws in place specifically targeting and discriminating a specific group and de facto is controlled by actions of the people.

Segregation can also occur on many different bases, with ethnic and economic segregation being the most prevalent in today’s society. The issue of multiculturalism is currently a topic of discussion with the rise of xenophobic political parties and groups throughout Sweden. Given the immigration policies in action during the past decades, Sweden has become more and more multicultural but the vastness of it all has become troublesome in creating an integrated society. Different municipalities accept very varied amounts of immigrants and this along with the fact that immigrants are usually placed in areas with cheaper accommodation, create segregation both between and within municipalities. The economic segregation has

---

1 Linköping University. *Institute for Analytical Sociology.* URL: http://www.liu.se/ias?l=en%5C&sc=true (visited on 05/18/2015)

been brought to the attention of the public in a very extreme way, with the influx of beggars from Eastern Europe and their complete isolation from the rest of society.

A large problem within the area of segregation are reports showing an increase in de facto segregation in the past couple of decades.\textsuperscript{3} One aspect of this problem is the segregation at workplaces, where ethnic discrimination in the hiring process is a big issue. Identifying what causes these patterns of segregation have been previously studied using social, as well as simple mathematical and statistical models, however with the level of technology and size of data present today, new methods should be investigated to see whether more information can be derived without completely exhausting resources at the same time. The data mining methodology is well suited for this task as the goal of these methods are to procure information, fast and simple, not based on any previous theories or knowledge from large amounts of data. Introducing the concept of frequent sequence mining on the longitudinal\textsuperscript{4} data present in Statistics Sweden’s Longitudinal integration database for health insurance and labor market studies (LISA)\textsuperscript{5} would allow for greater usage of the data as a whole. One efficient algorithm is called constrained Sequential PAttern Discovery using Equivalence classes (cSPADE) which is the algorithm of choice in this thesis.

This discussion gives rise to the following questions: How does ethnic discrimination affect the workplace and movement patterns? Do these xenophobic opinions present themselves where an individual works? This thesis will mainly explore ethnic segregation and the impact of the overall ethnic structure in a workplace on Swedish employees and the movement patterns of these individuals. The definition of Swedish is defined as an individual born in Sweden to Swedish-born parents.

1.3. Objective

The objective of this thesis is to investigate the effectiveness of a frequent sequence mining method, cSPADE, on longitudinal data at an individual level. The focus is finding patterns that explain the movement flow of Swedish employees to more segregated workplaces. The target group is Swedish employees at workplaces with between 10 and 100 employees. These limits are chosen assuming that the individual should be able to notice a change in ethnic diversity, which in larger workplaces would be harder. The lower limit is set as to not allow a single person having a too large of impact on the composition of the workplace. As the size of the data is so large, the target group data is split into two distinct categories with the motivation of reducing the process time of the algorithm.

\textsuperscript{3}Migrationsinfo. Segregation. 2013. URL: http://www.migrationsinfo.se/valfard/boende/segregation/ (visited on 05/09/2015)
\textsuperscript{4}A longitudinal data set is where the same individual is measured at different points in time.
\textsuperscript{5}By the Swedish acronym.
1.4 Related work

Previous studies on the topic of segregation have been focused on neighborhoods and traditional social models such as the Schelling model first described in 1971. He created the model as a way of describing the segregation of color in the United States. Schelling leaves out two main causes for segregation, organized action and the economic process, and only focuses on individual behavior. Schelling uses a one dimensional spatial neighborhood and creates a bias of the ratio of similar neighbors needed for happiness to simulate how a low individual bias aggregates into a larger cumulative bias for the entire population.

One of the findings in the simulations is that when the desire for similar neighbors lies around a third, the level of segregation is quite small, but when the desire increases to approximately half of its neighbors being similar, the level of segregation increases rapidly. Another model presented is the bounded-neighborhood model which describes another definition of a neighborhood. Instead of having a line or plane, a bounded area is defined as the neighborhood and an individual can either be inside or outside it.6

This opens up for analysis of other areas such as workplaces in a sense that an employee either works at a company or does not. These types of models have been used in previous studies on schools and childrens’ movement patterns to analyze the increasing segregation found in Swedish schools.7 Segregation in workplaces have been analyzed more focused on ethnic discrimination and how it affects the hiring process where multiple correspondence tests using Swedish sounding and foreign sounding names.8

The number of employees have been found to impact the level of segregation quite heavily as a smaller company tends to be more negative towards non-Swedish individuals.9 These models are not chosen as the method in this thesis as the objective is to see whether a specific data mining method can identify unknown information. Using the frequent sequence mining methods on longitudinal data has been studied previously with several proposed schemes of how this can be made. However these are not compatible with the implementation of the algorithm used in this thesis.10

---

10 Aída Jiménez, Fernando Berzal, and Juan Carlos Cubero. “Mining patterns from longitudinal studies”. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artifi-
Another data mining method called frequent pattern mining also exists but since it is of interest to find patterns that lead to a change rather than just measuring the co-occurrence of events for the individuals, the frequent sequence mining methodology is the focus of the thesis.
2. Data

2.1. Data sources

The data used in this analysis are mainly from Statistics Sweden’s LISA database. This database contains longitudinal information of the Swedish population aged 16 or older registered in Sweden on December 31st of each year between 1990 and 2012. Information found in this database relates to demographic, employment, and income information on the individual as well as geographic, demographic and economical information about companies and places of employment. Information from LISA is merged with various other databases from Statistics Sweden, described in the following section.

2.1.1. Databases

This section gives a brief description of the different databases used. Further explanations of the variables and their use in this thesis is explained in section 2.2.

LISA database

Each row in the data set consists of characteristics of one individual during one year. The database has two different groups of measurements defining the information about the individual’s employment measured at two different time points. One of these groups is recorded on the 30th of November every year while the other group records information about the largest source of income during the entire year meaning that an individual could have two different employers in these variables. In this thesis, the employer of the individual is defined as workplace connected to the largest source of income, thereby counting this as the only workplace during the whole year. This database uses two different serial numbers, one to identify an individual and one to identify a workplace. These are used in section 2.2 to merge information from the other databases.

Multiple generations database

Each row of this database contains information about the family of an individual, specifically the birth country of the parents and/or adoptive parents.

Background database

This database contains static information about the individual, i.e. information that does not change over time. Information taken from here is the birth country, gender and birth year.

Company and workplace dynamic database

The *Company and workplace dynamic* (FAD) contain information about companies and workplaces. Each row consists of several serial numbers for the current and previous year as well as information on any changes that have been made concerning the specific workplace, such as a decommission or merging.

Perceived foreignness index

The perceived foreignness index is a database containing the average perceived cultural difference of 50 countries compared to Sweden. The perceived foreignness index is based on an Inglehart–Welzel cultural map taken from the fifth installment of the World Value Surveys (WVS) between 2005 and 2007. The WVS surveys began in 1981 and have been through several installments in 1990, 1995-1998, 2000-2001 and again in 2010-2012. During each installment new countries are added but the questions remain very similar; meaning that the change in values over time can be tracked and visualized.

As can be seen in Figure 2.1, Inglehart et.al. present the results of the surveys in two dimensions: a traditional vs. secular-rational value and a survival vs. self-expression dimension. These were seen as the two major influences in determining the cultural and societal status of a country as they in all installments of the survey were deemed to explain more than half of the so called cross-national variation in the responses. The variance within nations is much smaller than the cross-national differences generating the conclusion that the nation, even in this time of globalization, is an important predictor of an individual’s values.

One must take into account that the perceived foreignness index is just the perceived difference between two individuals from different nationalities. In practice a person of a specific national heritage living in another country might have very different values compared to its nation as a whole.

The index is created by the Euclidean distance between Sweden and the other countries in Figure 2.1. For instance the perceived foreignness index of a Norwegian would be 0.456 while the perceived index of a Pakistani would be 3.034.\footnote{Ronald Inglehart and Christian Welzel. “Changing Mass Priorities: The Link between Mod-}

\footnote{By the Swedish acronym.}

\footnote{Ronald Inglehart and Christian Welzel. “Changing Mass Priorities: The Link between Mod-}
2.2 Data management

This section describes the variables used in the analysis. The first three variables are only row identifiers which are needed for the frequent sequence mining algorithm. The description also gives information on how many categories are present in each variable. In total 12 explanatory variables and 40 categories are given to the analysis. This section explains the creation of these variables.

As stated in section 2.1.1, two keys, the individual and workplace serial number, are used to merge information in the different databases. The data concerning the individual uses the individual serial number while all data concerning the workplaces....

Figure 2.1.: Ingelhart-Welzel cultural map of 2008 (Source: World Value Survey. World Value Surveys. URL: http://www.worldvaluessurvey.org/WVSContents.jsp (visited on 05/09/2015))
uses the workplace serial number. The last four variables presented in this section are discretized using the steps given in section 2.2.1. *Age*, *ethnic ratio* and *workplace size* are discretized using specific intervals.

**Individual serial number** - Variable identifying which individual is in the row.

**Time** - The time variable is created from the current year, giving the year 1990 a starting value of 1 to the last value of 23 representing 2012.

**Size** - This is a variable needed for the cSPADE algorithm to indicate how many variables are connected to each specific row. As there are always 12 variables for each row this number will always be 12.

**Education** - The education level is according to the Swedish Educational Terminology (SUN) 2000 standard; a three digit code providing the level, length and area of study. This thesis is focused on the level of education so only the first digit is used when constructing three different categories of education: pre-high school, university less than two years, and university greater than two years.\(^{14}\)

**Workplace size** - The most important addition to the data is counting the number of individuals working at each of the workplaces for each year to distinguish the different workplace sizes. The intervals of this variable are taken from the Statistics Sweden definition of workplace sizes, but merging the groups that are not the focus of the thesis, into the following categories: less than 10, 10 to 49, 50 to 99, 100 or more employees and a specific size category for unemployment.

**Unemployment** - If the previous year’s workplace serial number is missing then the individual is deemed to have been unemployed, otherwise not.

**Age** - The age is measured in years since birth, discretized into intervals of ten years.

**Ethnic ratio** - The ratio of non-Swedish employees is calculated by the number of individuals not fitting the Swedish definition, stated in section 1.2, divided by the total number of employees at the workplace. The discretized groups for this variable are given in intervals of size 0.2.

**Movement pattern** - In order to determine whether an individual has changed workplace between years, the workplace identifiers from the current and previous year are compared. As there exists a possibility that the workplace identifier for a workplace changes from one year to the next on account of administrative changes while no physical change is made, the FAD database is used to link workplaces with identifiers from both years. The different patterns created are: move to a new workplace with higher ethnic ratio, move to a new workplace with lower ethnic ratio, stay at the old workplace having a higher ethnic ratio, stay at the old workplace having a lower ethnic ratio.

**Move** - The place of residence and number of moves is used to create a new variable

---

combining these two pieces of information describing three different situations: a move within or outside the municipality or no move at all.

**Gender** - This variable is directly taken from the raw data, grouping the individual into the two different genders.

**Ind. income** - The income is taken from the LISA database as the largest source of income during the year measured in SEK.

**Ind. income change** - The change of income is created as a fraction of the individual’s income the current year divided by the previous.

**Perceived foreignness index** - The perceived foreignness workplace average is calculated by averaging the individual’s values working at each specific workplace each year.

**Workplace income** - This variable is created by averaging the individual’s income at each workplace each year.

### 2.2.1. Discretization

One of the criteria for frequent sequence mining is that the variables must be categorical. This creates a problem specifically for variables such as the income of an individual as it is provided as a yearly sum. In order to solve this issue, the numerical variables present in the final data set are discretized into a specific number of groups. This discretization is done by first creating a class for the missing values, then a class for the value 0 and finally calculating the 20th, 40th, 60th and 80th percentiles of every variable and creating a category with these percentiles as borders. This means that each numerical value is converted into seven categories. The value 0 is given its own class because of the large prevalence of 0 in the different numerical variables, for example unemployed individuals having an income of 0 and Swedish individuals having a perceived foreignness index of 0. Variables not containing any missing values are only converted into six categories.

### 2.2.2. Sub-setting the data

As the thesis is focused on workplaces with between 10 and 100 employees the first step is to limit the analyzed individuals to those who have worked at a workplace of this size at any time during the 23 years the data cover. In order for an individual to be used in the analysis, they also need to be recorded in every year of the data set. This is because the reasons for entering or leaving the data are many, deaths, immigration and emigration, so this simplifies the data set. There are 2 536 235 individuals that fulfill these two criteria. This data of individuals is split up into two samples by a random sample of 50 percent of the individuals because, as stated in section 1.3, the algorithm would take too long if the entire data set is used.
3. Methods

The following section gives a description of the methods used in the analysis.

3.1. Frequent sequence mining (FSM)

First follows an example used throughout the section to easily explain the methods, next a brief description of terminology used in the section and finally a description of the methods used in this thesis.

**Example**  Table 3.1 contains an example data set used in sequence analyses, which will be used throughout section 3.1 to further explain the different algorithms, terminologies and steps presented.

This data consist of four individuals, identified by the *sequence-id*, with measurements of events occurring at three different points in time, identified by the *event-id*. The column *events* contain the different events occurring for the specified individual at the specified time.

<table>
<thead>
<tr>
<th>sequence-id</th>
<th>event-id</th>
<th>events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>&lt;A, C, E&gt;</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>&lt;A, B, E&gt;</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>&lt;A, D, E&gt;</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>&lt;A, D, E&gt;</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>&lt;A, B, C&gt;</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>&lt;B, C, D&gt;</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>&lt;B, C, E&gt;</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>&lt;C, D, E&gt;</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>&lt;B, C, D&gt;</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>&lt;A, D, E&gt;</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>&lt;A, B, D&gt;</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>&lt;C, D, E&gt;</td>
</tr>
</tbody>
</table>
**Sequence**  Assume that \( I = \{A, B, C, D, E\} \) is the complete set of all the five distinct events present in the data seen in Table 3.1. The events could be anything from a purchase of a specific item, a recorded change of state of a switch in a system or the yearly income of a person. A sequence consists of an ordered set of elements written as \( S = < e_1 \rightarrow e_2 \rightarrow ... \rightarrow e_n > \), where \( e_i = \{i_1, i_2, ..., i_k\} \) is a collection of \( k \) events taken from \( I \) and \( e_j \) happens after \( e_i \) if \( j > i \). The size of the sequence is defined by the total number of events in the sequence. An observed sequence from the example is

\[
s = < A \rightarrow B \rightarrow D >
\]

which can be seen in individual 1, 2 and 4.

**Sub-sequence**  A sequence, \( T = < t_1 \rightarrow t_2 \rightarrow ... \rightarrow t_m > \), is a sub-sequence to another sequence if there exists an injective function that preserves the order of events and all \( t_i \subseteq e_j \). In the observed example above the sequence

\[
t = < A \rightarrow B >
\]

is a sub-sequence of \( s \) because the order is preserved and all elements of \( t \) are subsets of the corresponding elements in \( s \).

**Contiguous sub-sequence**  This type of sub-sequence is a special version needed for the use of different timing constraints, further explained in section 3.1.2.2. A sequence \( U \) is a contiguous sub-sequence of \( S \) if it fulfills any of these three criteria:

- \( U \) is created from \( S \) by removing an event from either \( e_1 \) or \( e_n \).
- \( U \) is created from \( S \) by removing an event from any \( e_i \in S \) that has more than one event.
- \( U \) is a contiguous sub-sequence of another sequence \( V \) which in turn is a contiguous sub-sequence of \( S \).

In the example shown, \( t \) is a contiguous sub-sequence of \( s \) as an event from the last element is removed in creating \( t \). \(^{15}\)

**Support**  The support value of sequence \( S \), denoted by \( \sigma(S) \), is seen as the number of sequences in the database containing \( S \). Often the support is written as \( \text{supp}(s) = \frac{\sigma(S)}{N} \), where \( N \) is the total number of sequences in the data, resulting in a proportion of the sequences containing \( S \). This measure is used when generating frequent sequences by comparing the support for every sequence to a threshold

3.1 Frequent sequence mining (FSM)

value, removing all sequences with a support lower than this value. In the example above \( \sigma(s) = 3 \) which gives a support of 0.75, i.e. the sequence \( s \) occurs in 75 percent of the data. A special property is present in the support, namely:

\[
\sigma(\text{sequence}) \leq \min(\sigma(\text{sub-sequence}))
\]

which means that any sequence cannot have a support value higher than the minimum support of any of its sub-sequences.

**Confidence** Confidence is one measure of how interesting a sequence is and is based on the support of a sequence compared to the support of the sub-sequence. In the example above the sequence \( s \) can be seen as the joint occurrence of the sub-sequence \( t \) and the event \( D \). In other words the confidence of \( s \) is calculated as

\[
\text{conf}(s) = \frac{\sigma(< t \rightarrow D >)}{\sigma(t)} = \frac{3}{3} = 1
\]

The result of this calculation is interpreted as event \( D \) happens in 100 percent of the time when \( t \) occurs.

**Lift** Lift is another measure of determining sequences of interest in the analysis comparing how many more times the observed sequence occurs compared to the occurrence if the events were independent. Considering the example above the calculation is

\[
\frac{\text{supp}(s)}{\text{supp}(t) \times \text{supp}(D)} = \frac{0.75}{0.75 \times 1} = \frac{1}{1} = 1
\]

This means that the probability of event \( D \) given the sequence \( t \) is equally likely as if \( t \) does not occur, meaning that the two sets of events are independent. Lift values higher than 1 means that the sequence is positively correlated, more likely to occur given the events compared to if the events not occur, and lift values lower than 1 means that the sequence is negatively correlated, less likely to occur.\(^\text{16}\)

**Lattice** A lattice is a partially ordered set of elements where you have a unique least upper bound and greatest lower bound. The lattices described further on in this section are ordered by the subset function which means that all sequences connected to a sequence on the level above in the graph seen in Figure 3.1 are sub-sequences of that sequence.\(^\text{17}\)

\(^\text{16}\)Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*

3.1.1. Apriori property

The apriori property is a very useful tool when generating frequent item-sets or sequences. Using a brute-force method of creating sequences, even from a data set only containing five distinct events, the total number of sequences quickly becomes very large. An example of the different sequences that can be produced can be seen in Figure 3.1. As can be seen, the different combinations of joining two single events are many, either joining them in the same or different elements. Another aspect of sequences is the possibility of repeating the same event in consequent elements. The figure shows a small subset of all the possible combinations of nodes in the graph, an example being that the sequence $< A >$ joined with $< B >$ forms the sequence $< A, B >$ as well as $< A \rightarrow B >$.\(^{18}\)

The apriori property states that if a specific sequence is frequent then all of its sub-sequences are also frequent. This in turn also means that if a specific sequence is infrequent then all of its super-sets are also infrequent. What this means in the context of generating frequent sequences, is that it is very easy to reduce the search space of possible candidate sequences in order to reduce the computational load compared to the brute-force method. Given a specific support threshold and the fact that

\[
\sigma(\text{sequence}) \leq \min(\sigma(\text{sub-sequence}))
\]

many parts of the sequence lattice are removed from the search. In Figure 3.1 for instance if the sequence $< A >$ does not meet the threshold, then according to the apriori property, no sequence containing that event in any element is frequent, i.e. all the nodes with multiple elements shown, thereby never wasting computational power generating and checking these sequences.\(^{19}\)

3.1.2. Method and algorithms

FSM is a method used on temporal data when trying to identify sequences that co-occur often. The type of data that is most often used with this method is market basket data which consist of information about the customer’s purchase behavior over time.\(^{20}\) This method focuses on categorical, or event, patterns rather than numerical data which means that any numerical variable needs to be discretized. There exist many different algorithms meant to efficiently generate frequent sequences in a transactional database, many of them incorporate something called the apriori property, further explained in section 3.1.1. Because the data tend to be very large, the chosen algorithm needs to be able to perform well and efficiently for both small and large data sets, i.e. it should be able to scale well.

---

\(^{18}\)Zaki, “SPADE: An efficient algorithm for mining frequent sequences”; Tan, Seinbach, and Kumar, *Introduction to Data Mining*

\(^{19}\)Tan, Seinbach, and Kumar, *Introduction to Data Mining*

\(^{20}\)ibid.
3.1 Frequent sequence mining (FSM)

One of the existing scalable algorithms is SPADE which was first presented in 1997. This algorithm has an implementation in the R-package `arulesSequences` and comparing the performance of SPADE versus two other algorithms GSP and PrefixSpan, it can be seen that PrefixSpan scales the best with SPADE being a close second.

### 3.1.2.1. SPADE

The main approach of SPADE is to reduce the sequence lattice into closed sub-lattices which are run independently from one another. The sequences are described by a vertical id-list that defines in which object and time the events occur. This means that the database is only scanned at most three times compared to once for each different length of the sequence in other algorithms. SPADE also works with a specific data format, an example shown in Table 3.1, with the assumption that no sequence has any duplicate observations.

The first step of the algorithm is to produce sequences of size one and scanning the database for their frequency, also called support. The definition of a frequent

---


sequence is determined by a support threshold, i.e., the least amount of times a sequence has to occur in the data in order to be seen as interesting and not random. The support acts as a constraint in the algorithm, producing only sequences that have a large presence in the data. The frequencies of size one that meet the threshold are then used to form sequences of size two by a simple join on the id-list. So far the algorithm has scanned the database two times, one for each sequence size, and saved the id-lists of the frequent sequences of size one and two in the main memory.

The next step is to produce all remaining frequent sequences. Firstly the full sequence lattice is split into several sub-lattices. Next the frequent \( k-1 \) sequences are seen as singular building blocks and are joined together in all possible combinations while checking the support of the resulting sequences. For instance when creating the sequences of size 3 in Figure 3.1, all sequences of size 2 are seen as single building blocks and are combined in every possible combination.\(^{23}\)

### 3.1.2.2. cSPADE

cSPADE is a modification to the aforementioned SPADE algorithm that allows for additional constraints. These are constraints meant to reduce the search space of the algorithm to both reduce the resources needed to produce the sequences and to focus the algorithm in producing interesting sequences. There are three different time constraints implemented in this algorithm:

- **mingap**, defining the minimum time allowed between consecutive elements in a sequence.
- **maxgap**, defining the maximum time allowed between consecutive elements in a sequence.
- **maxwin**, defining the maximum time between first and last element in a sequence.
- **maxsize**, defining the maximum number of events allowed in an element.
- **maxlen**, defining the maximum number of elements in a sequence.

Using the timing constraints a modification of the apriori property is needed in order for the algorithm to not mistakenly generate infrequent sequences. The property then states that if a sequence is frequent then all contiguous sub-sequences must also be frequent.\(^{24}\)

\(^{23}\)Zaki, “SPADE: An efficient algorithm for mining frequent sequences”


20
3.1 Frequent sequence mining (FSM)

3.1.3. Pruning frequent sequences

Sequences that meet both the constraint criteria from the sequence generation and a confidence threshold might or might not be deemed interesting and the next step in the FSM procedure is pruning them to find the most interesting. A simple way to do this is to sort the confidence and/or lift measurements in descending order. Usually sequences with a high confidence and lift indicate that the events are occurring together by something more than pure randomness.

There is also another way to prune sequences when it comes to explaining different classes. If the intention of the FSM is to find patterns that explain one specific group of events, sequences explaining the desired event can be compared to sequences explaining the non-desired event(s) to see if there are any unique patterns. This comparison is done by matching the sequences for both classes and comparing the different measurements, with the assumption that those that occur often in both classes are not deemed good enough as a predictive pattern for the desired class.

The next step in the pruning is to remove redundant patterns, i.e. sequences that have a confidence equal or lower than that of a sub-sequence, by the assumption that the sub-sequence is deemed to have an equal or better predictive power with less information. Given the example in the beginning of this section, the sequence $< AB \rightarrow C >$ has a confidence of $\frac{2}{3} = 0.67$. This sequence is then compared to all the sub-sequences and checks the confidence. The confidence of the sub-sequence $< B \rightarrow C >$ is $\frac{3}{4} = 0.75$ meaning that the sub-sequence is better at predicting the resulting event, making $< AB \rightarrow C >$ redundant.

---

4. Results

The following sections present the results from the data processing and frequent sequence mining.

4.1. Data visualization

After the data went through the processing described in section 2.2, the class distribution of the variables are interesting to visualize further.

Figure 4.1 shows the size distribution of workplaces, clearly indicating that the most frequent size category is between 10 and 49 employees with approximately 30 percent of the individuals working at these workplaces. The size more than 100 have approximately 23 percent of the observations but the sizes in this group vary from 100 to 10000.

![Distribution of the workplace sizes](image)

**Figure 4.1.:** Distribution of the workplace sizes

Figure 4.2 shows the distribution of the class variable being the focus of this analysis, the movement pattern. As only approximately nine percent of the patterns is a move of the individual to a new workplace with a lower ethnic ratio, i.e. more segregated, it can be expected that there will not be as many frequent sequences explaining this class given that almost 45 percent describes the opposite event, a move to a new workplace with a higher ethnic ratio. The class *HigherOLD* defines a person who has
not moved to a new workplace but the ethnic ratio is higher while the class \textit{NONE} indicate that the information is from the 1990 record for every individual meaning that there is no information from 1989 to compare with. The class \textit{LowerOLD} is one class that does not exist at all in the data meaning that the ethnic ratio of the same workplace always increased.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig4_2}
\caption{Distributions of movement pattern in four classes}
\end{figure}

As the analysis is focused on workplaces of specific sizes, Figure 4.3 shows the distribution of the movement patterns compared to the workplace sizes. Comparing the two analyzed classes, it can be seen that the relative distribution of workplaces between 10 and 100 employees are approximately equal. An interesting note from this figure is that staying at the old workplace with a higher ethnic ratio is in almost 30 percent consisting of unemployed individuals.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig4_3}
\caption{Grouped distribution of the workplace sizes within each movement pattern}
\end{figure}
4.2 Sequence mining

The ethnic ratio of the workplace is shown in Figure 4.4. The majority of all workplaces have a very low ratio of non-Swedish employees, with only about two percent of all workplaces having a majority of non-Swedish employees.

Figure 4.4.: Distribution of the ethnic ratio

4.2. Sequence mining

As the focus of this thesis is looking at the movement patterns of Swedish employees, the sequences that are primarily analyzed from the different runs of the algorithm are the ones describing an individual changing workplaces to either a higher or lower ethnic ratio. In the following sections, the ten sequences with the highest support and confidence will be provided and analyzed. Because the focus is Swedish employees of workplaces of a specific size, only sequences containing the two size groups, from 10 to 49 or 50 to 99 employees, in the left hand side are included.

Given the apriori property, as the support threshold is lowered, more sequences are deemed frequent and the search lattice expands quite rapidly. This means that the time it takes for the algorithm to mine all frequent sequences increases. The following sections investigate how different support thresholds affect the processing time for the algorithm and the resulting sequences.

4.2.1. Multiple events in an element

In this section, the following constraints were used:

- A support threshold of 50 and 40 percent.
- Maximum number of events in an element, $maxsize = 2$.
- Maximum number of elements in a sequence, $maxlen = 3$. 
• Minimum time between consecutive elements in a sequence, \( \text{mingap} = 1 \).
• Maximum time between consecutive elements in a sequence, \( \text{maxgap} = 2 \).
• Maximum time between first and last element in a sequence, \( \text{maxwin} = 5 \).
• Confidence threshold of 50 percent.

What this means in practice is that the sequences are allowed to have a maximum of three elements with a maximum of two concurrent events. Consecutive elements can at most occur within two years and the whole sequence can at most envelop a five year window. These settings were chosen so that an event happening one year would not be influenced by an event more than five years prior.

The time it took for the algorithm to mine all frequent sequences meeting the support threshold of 50 percent was approximately 28 hours. The algorithm produced 285307 sequences, of which only 22 are sequences explaining a move to a workplace with a lower ratio.

The 10 sequences found with the highest confidence and support, seen in Table 4.1, are not very informative or interesting. They only contain information of one workplace size, that the individual does not move and has been employed the previous year. Comparing these sequences to the ones shown in Table 4.2, it is clear that these events also explain, with a higher support, confidence and lift, a move to a workplace with a higher ratio. The same result holds for all of the 22 sequences found which means that there are no unique sequences that can explain why a Swedish employee moves to a workplace with a lower ethnic ratio.

An interesting note from both tables is that the lift value for all sequences are lower than 1 meaning that the events in the left hand side of the rule have a negative correlation with the movement pattern. What this means is that it is less likely to move to a new workplace given the stated events compared to if the stated events do not occur.

Lowering the support threshold to 40 percent, the algorithm took approximately 55 hours to find all frequent sequences. From the resulting 755637 frequent sequences, only 312 sequences explain the desired class. However the same problem is present here, none of the 312 sequences have a higher predictive power than sequences with the same events describing a move to a workplace with a higher ethnic ratio.

4.2.2. Single event in an element

As stated earlier, the lower the support threshold, the more sequences are deemed frequent increasing the search lattice and the processing time. As the 40 percentage support threshold already took more than 50 hours to finish, a lower threshold with the same constraints led to an unsustainable process time. The decision was then made to lower the \( \text{maxsize} \) constraint to 1 which resulted in much quicker processing time however restricting the simultaneous occurrence of events in an element to only
1. With a support level of 30 percent it took approximately one and a half hours to generate all the frequent sequences. The benchmarks for the algorithm with different levels of support can be seen in Table 4.3. The table shows that the processed time does not increase as quickly when the support threshold is decreased compared to in section 4.2.1. Notably the amount of frequent sequences are only one tenth or less of the number compared to the previous runs.

These sequences provide more interesting information, for instance Table 4.4 indicates that

\(< \{\text{Size} = 10 \rightarrow 49\}, \{\text{Ratio} = -0.2\} \rightarrow < \{\text{Ratio} = \text{Lower NEW}\} >\)

occurs in 47.2 percent of all sequences, meaning an individual moves to a new workplace with a lower ethnic ratio if the ratio of the old workplace is less than 20 percent. Similar to the previous analyses, when comparing with the sequences from Table 4.5, no unique sequences that indicate a move to a workplace with a lower ethnic ratio can be found in neither of the three different support thresholds.

An interesting note from both tables is that the lift value for all sequences are lower than 1 meaning that the events in the left hand side of the rule have a negative correlation with the movement pattern. What this means is that it is less likely to move to a new workplace given the stated events compared to if the stated events do not occur.

### Table 4.3.: Benchmarks of different support thresholds

<table>
<thead>
<tr>
<th>Support</th>
<th>Process time (hours)</th>
<th>Frequent sequences</th>
<th>Desired class</th>
<th>Unique sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>1.4</td>
<td>23337</td>
<td>37</td>
<td>No</td>
</tr>
<tr>
<td>25</td>
<td>1.7</td>
<td>28981</td>
<td>56</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>2.2</td>
<td>46613</td>
<td>116</td>
<td>No</td>
</tr>
</tbody>
</table>

### 4.2.3. General results

Looking into the other sequences from the algorithm the first step is removing redundant sequences. Given the number of sequences present in section 4.2.1, pruning these takes too long time so the analysis is done on the results from section 4.2.2 with a 30 percent support threshold.

From the original 23337 frequent sequences the pruning of redundant sequences, i.e. sequences of which a sub-sequence has a better predictive power, results in only 1659 sequences remaining. After removing non-interesting sequences, e.g. a pattern of gender leading to gender and similar non-descriptive patterns, only 1161 sequences remain. These sequences are then ordered with respect to the three measures of interest, support, confidence and lift. The top 20 sequences from each sorting are available in the appendix for further details.
Table A.1 specify sequences that have a high occurrence in the data. Almost half of the 1161 sequences occur in the majority of the data, i.e. a support of more than 50 percent, of its neighbors need to be similar with varied values of confidence and lift. Many of these sequences explain movement patterns based on specific income changes, ethnic ratio or previous movement patterns. Looking at Table A.2, the sequences with higher confidence explain either a movement pattern, a change in income of more than 135 percent or an ethnic ratio of less than 20 percent. When sorted by lift most of the top sequences concern individuals being unemployed, identified by the individual or workplace income being 0. Many positively correlated sequences also contain patterns to a workplace with a perceived foreignness index of more than 0.342.
### Table 4.1: Most interesting sequences leading to lower ethnic ratio with 50 percent support threshold

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (&lt;{\text{Size}=10-49}, {\text{Move}=\text{No}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.549</td>
<td>0.640</td>
<td>0.778</td>
</tr>
<tr>
<td>2 (&lt;{\text{Size}=10-49}, {\text{FromEmployed}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.549</td>
<td>0.638</td>
<td>0.777</td>
</tr>
<tr>
<td>3 (&lt;{\text{Size}=10-49}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.546</td>
<td>0.629</td>
<td>0.766</td>
</tr>
<tr>
<td>4 (&lt;{\text{Size}=10-49}, {\text{FromEmployed, Move}=\text{No}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.537</td>
<td>0.628</td>
<td>0.764</td>
</tr>
<tr>
<td>5 (&lt;{\text{FromEmployed, Size}=10-49}, {\text{Move}=\text{No}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.527</td>
<td>0.631</td>
<td>0.768</td>
</tr>
<tr>
<td>6 (&lt;{\text{FromEmployed, Size}=10-49}, {\text{FromEmployed}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.527</td>
<td>0.630</td>
<td>0.767</td>
</tr>
<tr>
<td>7 (&lt;{\text{Move}=\text{No, Size}=10-49}, {\text{FromEmployed}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.525</td>
<td>0.625</td>
<td>0.761</td>
</tr>
<tr>
<td>8 (&lt;{\text{Move}=\text{No, Size}=10-49}, {\text{Move}=\text{No}}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.525</td>
<td>0.626</td>
<td>0.762</td>
</tr>
<tr>
<td>9 (&lt;{\text{FromEmployed, Size}=10-49}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.523</td>
<td>0.619</td>
<td>0.754</td>
</tr>
<tr>
<td>10 (&lt;{\text{Move}=\text{No, Size}=10-49}&gt; \implies {\text{Ratio}=\text{LowerNEW}}&gt;)</td>
<td>0.521</td>
<td>0.613</td>
<td>0.746</td>
</tr>
</tbody>
</table>
### Table 4.2: Most interesting sequences leading to higher ethnic ratio with 50 percent support threshold

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>${\text{Size}=10-49} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.831</td>
<td>0.957</td>
<td>0.960</td>
</tr>
<tr>
<td>${\text{Size}=10-49} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.825</td>
<td>0.950</td>
<td>0.958</td>
</tr>
<tr>
<td>${\text{Size}=10-49} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.820</td>
<td>0.945</td>
<td>0.950</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{Move}=\text{No}} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.819</td>
<td>0.953</td>
<td>0.956</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{FromEmployed}} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.817</td>
<td>0.950</td>
<td>0.954</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{FromEmployed}, \text{Move}=\text{No}} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.812</td>
<td>0.949</td>
<td>0.952</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{Move}=\text{No}} &lt; {\text{Move}=\text{No}, \text{Ratio}=\text{HigherNEW}}$</td>
<td>0.811</td>
<td>0.945</td>
<td>0.950</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{FromEmployed}} &lt; {\text{Move}=\text{No}, \text{Ratio}=\text{HigherNEW}}$</td>
<td>0.810</td>
<td>0.942</td>
<td>0.947</td>
</tr>
<tr>
<td>${\text{Move}=\text{No}, \text{Size}=10-49} &lt; {\text{Ratio}=\text{HigherNEW}}$</td>
<td>0.810</td>
<td>0.954</td>
<td>0.958</td>
</tr>
<tr>
<td>${\text{Size}=10-49}, {\text{FromEmployed}, \text{Move}=\text{No}} &lt; {\text{FromEmployed}, \text{Ratio}=\text{HigherNEW}}$</td>
<td>0.809</td>
<td>0.942</td>
<td>0.950</td>
</tr>
</tbody>
</table>
### Table 4.4: Most interesting sequences leading to lower ethnic ratio with 30 percent support threshold

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><code>&lt;{Size=10-49},{Move=No}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.549</td>
<td>0.640</td>
</tr>
<tr>
<td>2</td>
<td><code>&lt;{Size=10-49},{FromEmployed}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.549</td>
<td>0.638</td>
</tr>
<tr>
<td>3</td>
<td><code>&lt;{Size=10-49}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.546</td>
<td>0.629</td>
</tr>
<tr>
<td>4</td>
<td><code>&lt;{Move=No},{Size=10-49}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.519</td>
<td>0.612</td>
</tr>
<tr>
<td>5</td>
<td><code>&lt;{FromEmployed},{Size=10-49}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.518</td>
<td>0.614</td>
</tr>
<tr>
<td>6</td>
<td><code>&lt;{Size=10-49},{Ratio=HigherNEW}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.498</td>
<td>0.600</td>
</tr>
<tr>
<td>7</td>
<td><code>&lt;{Size=10-49},{Ratio=HigherOLD}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.497</td>
<td>0.604</td>
</tr>
<tr>
<td>8</td>
<td><code>&lt;{Ratio=-0.2},{Size=10-49}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.487</td>
<td>0.601</td>
</tr>
<tr>
<td>9</td>
<td><code>&lt;{Size=10-49},{Ratio=-0.2}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.472</td>
<td>0.589</td>
</tr>
<tr>
<td>10</td>
<td><code>&lt;{Ratio=HigherNEW},{Size=10-49}&gt; =&gt; &lt;{Ratio=LowerNEW}&gt;</code></td>
<td>0.469</td>
<td>0.585</td>
</tr>
</tbody>
</table>
### Table 4.5: Most interesting sequences leading to higher ethnic ratio with 30 percent support threshold

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.831</td>
<td>0.957</td>
<td>0.960</td>
</tr>
<tr>
<td>2</td>
<td>0.819</td>
<td>0.953</td>
<td>0.956</td>
</tr>
<tr>
<td>3</td>
<td>0.817</td>
<td>0.950</td>
<td>0.954</td>
</tr>
<tr>
<td>4</td>
<td>0.808</td>
<td>0.953</td>
<td>0.957</td>
</tr>
<tr>
<td>5</td>
<td>0.805</td>
<td>0.954</td>
<td>0.957</td>
</tr>
<tr>
<td>6</td>
<td>0.786</td>
<td>0.945</td>
<td>0.949</td>
</tr>
<tr>
<td>7</td>
<td>0.772</td>
<td>0.953</td>
<td>0.956</td>
</tr>
<tr>
<td>8</td>
<td>0.767</td>
<td>0.932</td>
<td>0.935</td>
</tr>
<tr>
<td>9</td>
<td>0.766</td>
<td>0.958</td>
<td>0.961</td>
</tr>
<tr>
<td>10</td>
<td>0.766</td>
<td>0.954</td>
<td>0.958</td>
</tr>
</tbody>
</table>
5. Discussion

The objective of this thesis was to investigate the efficiency of the frequent sequence mining method cSPADE on a longitudinal data set of the population in Sweden from 1990 to 2012, specifically focusing on workplace movement patterns of Swedish employees. The results found from the analysis was that no unique patterns exist that would explain a Swedish employee moving to a new workplace with a higher segregation, i.e. lower ethnic ratio.

5.1. Evaluation of the results

The desired event for this analysis was a movement pattern to a new workplace with a lower ethnic ratio but the fact that this class was only present in only less than nine percent of the data presented some difficulties in finding sequences that would explain this move. One possible explanation for this is that the general ethnic ratio in Sweden increased over the years making it more probable that a workplace has an increase in ethnic ratio rather than a decrease. However not finding any unique sequences that would explain this movement pattern was not expected as well as the lift values of all the sequences being under 1. The conclusion that can be drawn from these findings is that the analyzed variables are not good at explaining the movement patterns present. This might be an effect of the difficulty of pinpointing specific reasons for a move of workplace, many of which are not present in the analyzed data and are difficult to measure, such as network dynamics and family structures. With this knowledge, looking at sequences with only one, perhaps even two, event(s) in an element would not reasonably provide a sufficient explanation of the move better than the pure randomness present in the data. The low lift value in all the analyses are indicators of the sequences being negatively correlated, i.e. the probability of moving to a new workplace is less given the events explaining the move compared to in general.

The variables that have the largest support and confidence for the two classes in the tables presented in section 4.2, are not that interesting or versatile. The majority of the sequences incorporate only information such as a move of residence or the individual’s employment status the previous year. Only a few sequences in Table 4.5 have events that incorporate the yearly earnings or the ethnic ratio of the workplace, but these sequences do not explain a move to a more segregated workplace.
Looking at the general results from the algorithm, it can be concluded that the sequences consist of many non-interesting patterns such as describing consistent unemployment or patterns of static variables. There also exist a lot of redundant sequences that can be removed in order to focus on the more important sequences found. The sequences that are located at the top when sorting by different measures consist of very similar variables, probably caused by the fact that each sequence from the original data is 23 elements long. This means that it is very probable that a specific sequence of length two or three can occur especially if the sequence contain events that are static throughout the data such as gender. This is why the lift measure is valuable in determining which of these sequences actually provide some information, alas in this case the lift values are mostly below 1 meaning that most sequences are negatively correlated and explain a pattern that decreases the probability of a certain event occurring. It is usually of interest to identify positively correlated sequences which can act as a base for finding reasons a specific event happens.

Is it reasonable to assume that the individual knows the ethnic ratio of a workplace before working there or even the current place of employment? In order to make this assumption more believable, the data is subsetted by reducing the sizes of workplaces analyzed to between 10 and 100 so that an individual is more probable to have a feel of the ethnic structure of the workplace. In this thesis, the workplace of the individual is defined as the primary source of income, meaning the workplace of an individual where the most income was earned. This definition is good at reducing the number of workplaces of an individual every year however it lacks the ability to account for how long the employment was at this specific workspace.

5.2. Evaluation of the methodology

Considering the complex nature of employment, there are many different reasons for switching jobs. Many of these reasons are aspects which are hard to measure such as networks of people and contacts as well as changes in the surroundings, e.g. family or company. The FSM analysis was conducted with many different events present but given the complexity many more or other variables could have been used. The biggest limitation of this method is that the computational complexity is exponentially connected to the amount of events, as the full search space lattice is created with all different combinations of events. However the complexity is also affected by the number of observations which is the biggest reason for the large process times in this thesis as the total number of sequences was approximately two and a half million. This leads to the conclusion that further sub-setting of the data is useful in reducing the process time needed for the algorithm.

The FSM needs discretized variables in order to see them as events, which means that a lot of information can be lost when converting continuous variables to discrete groups if the interval having the most effect is accidentally split by a boundary.
5.3 Further study

There exist a discretization process where as little information as possible is lost, by continuously joining neighboring groups starting with the smallest interval length possible. This method is a very extensive task as the analysis must be conducted on each different combination of interval lengths in order to see where the specific pattern actually occurs. The decision to skip this part in the thesis was due to the fact of the large amount of events already present in the data set producing very long processing times.

In the analyzed data, there exist variables that are constant for every year, such as gender, and instead of using the age of the individual, the year of birth would also be constant. They were included in the model with the hope of being a part of sequences that would explain specific movement patterns. These and other constant variables can instead be used in another method called multidimensional frequent sequence mining which subsets the analysis on constant variables. This would remove the need of having that variable present in the sequences and taking up one event in the sequences as well as sub-setting the data further. This would restrict the number of observations further and speeding up the algorithms but creating the need for an analysis done on every class of the variables.

5.3. Further study

The thesis has focused on specific variables and a specific way of discretizing continuous variables. It is possible to debate whether the variables used are more or less interesting or vital for the analysis while other variables that do exist in the different databases should or could have an impact on the movement patterns. Further study should be done on the variable importance of current and new variables in order to provide a better predictive quality. For instance, the structure or type of the workplace or family related variables might have a relation with the movement patterns. More time should also be spent on pruning the sequences from the variables that are deemed not important for the interpretation of the data.

One of the major issues with frequent sequence mining is that the results can only be seen as a co-occurrence of the specific sequence in the data rather than a causal relationship between the different elements of the sequence. This is why the support value is important in determining how frequent a sequence is, being a measure of the strength of the relationship. One way to tackle this problem is to conduct a confirmatory analysis, for example by the help of a multilevel model in order to measure and confirm the relationships found through the FSM. These models are meant to estimate data where variations are measured in multiple levels, for instance on the individual and workplace levels. Given the sub-setting of data that was conducted, the remaining 50 percent of the individuals can be used as a test group for any relations that might have been found from the FSM but instead creating a multilevel or other probabilistic model.
This thesis is based on the individual and its movement patterns. The same methods could be used but instead focusing on workplaces to see if there are any workplace characteristics or events that explain a change in ethnic ratio inside the specific workplace.

Since the data is measured over time it could also be represented as a time series. The ethnic ratio of a workplace could be analyzed using ARIMA modeling with other explanatory time series, also called transfer functions. This method would be useful in predicting the ratio of a workplace but would not be sufficient if explaining the data is of interest.
6. Conclusions

• There does not exist any patterns specifically defining why a Swedish employee moves to a new workplace with a lower ethnic ratio. All sequences describing a move to a new workplace are negatively correlated.

• Given the length of the sequences in the data, the algorithm produces many sequences containing and explaining static events resulting in many non-interesting sequences.

• In order for this method to be useful on this type of data, other variables should be either included or substituted into the analysis to see whether they have an effect, but at the same time restricting the amount of events present in the data.

• The cSPADE algorithm produce many patterns that are redundant, indicating that the algorithm needs to incorporate a pruning of these in a better way as the process time is very long.

• One way of lowering the process time of the algorithm further is sub-setting of the data, either by multidimensional FSM or by using a subset of the population.
A. Tables

A.1. General results
## Table A.1: Top 20 sequences sorted by support

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule=0.2</td>
<td>0.989</td>
<td>0.992</td>
<td>0.997</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.984</td>
<td>0.987</td>
<td>0.990</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.980</td>
<td>0.980</td>
<td>0.983</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.979</td>
<td>0.984</td>
<td>0.988</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.978</td>
<td>0.983</td>
<td>0.987</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.977</td>
<td>0.990</td>
<td>0.993</td>
</tr>
<tr>
<td>Ratio=-0.2</td>
<td>0.975</td>
<td>0.991</td>
<td>0.996</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.974</td>
<td>0.987</td>
<td>0.992</td>
</tr>
<tr>
<td>Ratio=-0.2</td>
<td>0.974</td>
<td>0.989</td>
<td>0.993</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.971</td>
<td>0.975</td>
<td>0.988</td>
</tr>
<tr>
<td>Ratio=-0.2</td>
<td>0.966</td>
<td>0.982</td>
<td>0.997</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.962</td>
<td>0.966</td>
<td>0.981</td>
</tr>
<tr>
<td>Ratio=-0.2</td>
<td>0.961</td>
<td>0.976</td>
<td>0.989</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.955</td>
<td>0.960</td>
<td>0.973</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.952</td>
<td>0.952</td>
<td>0.955</td>
</tr>
<tr>
<td>Rule=0.2</td>
<td>0.949</td>
<td>0.954</td>
<td>0.969</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.947</td>
<td>0.960</td>
<td>0.975</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.946</td>
<td>0.946</td>
<td>0.961</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.943</td>
<td>0.943</td>
<td>0.956</td>
</tr>
<tr>
<td>IncIndChange=1.351+</td>
<td>0.940</td>
<td>0.952</td>
<td>0.965</td>
</tr>
</tbody>
</table>
### Table A.2.: Top 20 sequences sorted by confidence

<table>
<thead>
<tr>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  &lt;{Gender=2}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.506</td>
<td>0.998</td>
<td>1.003</td>
</tr>
<tr>
<td>2  &lt;{Gender=1}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.491</td>
<td>0.997</td>
<td>1.001</td>
</tr>
<tr>
<td>3  &lt;{Gender=2}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.505</td>
<td>0.996</td>
<td>0.999</td>
</tr>
<tr>
<td>4  &lt;{Gender=1}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.489</td>
<td>0.992</td>
<td>0.997</td>
</tr>
<tr>
<td>5  &lt;{Ratio=HigherNEW}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.989</td>
<td>0.992</td>
<td>0.997</td>
</tr>
<tr>
<td>6  &lt;{SunLevel=gym-lt2}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.663</td>
<td>0.991</td>
<td>0.995</td>
</tr>
<tr>
<td>7  &lt;{Ratio=-0.2}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.975</td>
<td>0.991</td>
<td>0.996</td>
</tr>
<tr>
<td>8  &lt;{IncIndChange=1.351+}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.977</td>
<td>0.990</td>
<td>0.993</td>
</tr>
<tr>
<td>9  &lt;{Ratio=-0.2}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.974</td>
<td>0.989</td>
<td>0.993</td>
</tr>
<tr>
<td>10 &lt;{Gender=1}&gt; =&gt; &lt;{IncIndChange=1.351+}&gt;</td>
<td>0.487</td>
<td>0.988</td>
<td>1.002</td>
</tr>
<tr>
<td>11 &lt;{Age=26-35}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.569</td>
<td>0.988</td>
<td>0.991</td>
</tr>
<tr>
<td>12 &lt;{SunLevel=gym-lt2}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.661</td>
<td>0.987</td>
<td>0.992</td>
</tr>
<tr>
<td>13 &lt;{Ratio=HigherNEW}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.984</td>
<td>0.987</td>
<td>0.990</td>
</tr>
<tr>
<td>14 &lt;{IncIndChange=1.351+}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.974</td>
<td>0.987</td>
<td>0.992</td>
</tr>
<tr>
<td>15 &lt;{Gender=2}&gt; =&gt; &lt;{IncIndChange=1.351+}&gt;</td>
<td>0.500</td>
<td>0.985</td>
<td>0.998</td>
</tr>
<tr>
<td>16 &lt;{SunLevel=gym-lt2}&gt; =&gt; &lt;{IncIndChange=1.351+}&gt;</td>
<td>0.659</td>
<td>0.985</td>
<td>0.998</td>
</tr>
<tr>
<td>17 &lt;{Ratio=HigherOLD}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.979</td>
<td>0.984</td>
<td>0.988</td>
</tr>
<tr>
<td>18 &lt;{Gender=1}&gt; =&gt; &lt;{Ratio=-0.2}&gt;</td>
<td>0.484</td>
<td>0.983</td>
<td>0.999</td>
</tr>
<tr>
<td>19 &lt;{Ratio=HigherOLD}&gt; =&gt; &lt;{Ratio=HigherOLD}&gt;</td>
<td>0.978</td>
<td>0.983</td>
<td>0.987</td>
</tr>
<tr>
<td>20 &lt;{Age=-25}&gt; =&gt; &lt;{Ratio=HigherNEW}&gt;</td>
<td>0.304</td>
<td>0.982</td>
<td>0.985</td>
</tr>
<tr>
<td>Rule</td>
<td>Support</td>
<td>Confidence</td>
<td>Lift</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>------------</td>
<td>------</td>
</tr>
<tr>
<td>1. Size=Unemployed, IncWork=0</td>
<td>0.308</td>
<td>0.788</td>
<td>1.623</td>
</tr>
<tr>
<td>2. Size=Unemployed, IncInd=0</td>
<td>0.307</td>
<td>0.786</td>
<td>1.623</td>
</tr>
<tr>
<td>3. Ratio=0.2-0.4, IncWork=0</td>
<td>0.316</td>
<td>0.777</td>
<td>1.601</td>
</tr>
<tr>
<td>4. Ratio=0.2-0.4, IncInd=0</td>
<td>0.315</td>
<td>0.774</td>
<td>1.600</td>
</tr>
<tr>
<td>5. Ratio=HigherOLD, IncWork=0</td>
<td>0.335</td>
<td>0.766</td>
<td>1.579</td>
</tr>
<tr>
<td>6. Ratio=HigherOLD, IncInd=0</td>
<td>0.334</td>
<td>0.763</td>
<td>1.577</td>
</tr>
<tr>
<td>7. IncWork=0</td>
<td>0.364</td>
<td>0.753</td>
<td>1.551</td>
</tr>
<tr>
<td>8. IncInd=0</td>
<td>0.364</td>
<td>0.750</td>
<td>1.549</td>
</tr>
<tr>
<td>9. IncIndChange=0</td>
<td>0.353</td>
<td>0.749</td>
<td>1.547</td>
</tr>
<tr>
<td>10. IncInd=0</td>
<td>0.353</td>
<td>0.749</td>
<td>1.545</td>
</tr>
<tr>
<td>11. Size=-9, Size=-9</td>
<td>0.359</td>
<td>0.820</td>
<td>1.466</td>
</tr>
<tr>
<td>12. Size=Unemployed, CultDistWork=0.342+</td>
<td>0.300</td>
<td>0.685</td>
<td>1.416</td>
</tr>
<tr>
<td>13. Size=Unemployed, CultDistWork=0.342+</td>
<td>0.301</td>
<td>0.686</td>
<td>1.415</td>
</tr>
<tr>
<td>14. Ratio=HigherOLD, IncIndChange=Missing</td>
<td>0.305</td>
<td>0.674</td>
<td>1.393</td>
</tr>
<tr>
<td>15. Ratio=HigherOLD, IncIndChange=Missing</td>
<td>0.306</td>
<td>0.675</td>
<td>1.392</td>
</tr>
<tr>
<td>16. Size=Unemployed, IncIndChange=Missing</td>
<td>0.313</td>
<td>0.672</td>
<td>1.388</td>
</tr>
<tr>
<td>17. Ratio=0.2-0.4, IncIndChange=Missing</td>
<td>0.303</td>
<td>0.671</td>
<td>1.387</td>
</tr>
<tr>
<td>18. Size=Unemployed, IncIndChange=Missing</td>
<td>0.313</td>
<td>0.673</td>
<td>1.387</td>
</tr>
<tr>
<td>19. Ratio=0.2-0.4, IncInd=0</td>
<td>0.303</td>
<td>0.672</td>
<td>1.386</td>
</tr>
<tr>
<td>20. Size=100+, Size=100+</td>
<td>0.492</td>
<td>0.875</td>
<td>1.343</td>
</tr>
</tbody>
</table>
Bibliography


Jiménez, Aída, Fernando Berzal, and Juan Carlos Cubero. “Mining patterns from longitudinal studies”. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7121 LNAI.PART 2 (2011), pp. 166–179. ISSN: 03029743. DOI: 10.1007/978-3-642-25856-5_13.


