Linköping Studies in Science and Technology

Dissertations, No. 1331

Everyday mining

Exploring sequences in event-based data

Katerina Vrotsou

Linköping University
INSTITUTE OF TECHNOLOGY

Department of Science and Technology
Linköping University

Norrköping 2010

Everyday mining:
Exploring sequences in event-based data

Copyright © 2010 Katerina Vrotsou unless otherwise noted
katerina.vrotsou@itn.liu.se

Department of Science and Technology, Linköping University
SE-601 74 Norrköping, Sweden

ISBN 978-91-7393-343-8
ISSN 0345-7524

This thesis is available online through Linköping University Electronic Press:
www.ep.liu.se

Printed by LiU-Tryck, Linköping, Sweden

Abstract

Event-based data are encountered daily in many disciplines and are used for various purposes. They are collections of ordered sequences of events where each event has a start time and a duration. Examples of such data include medical records, internet surfing records, transaction records, industrial process or system control records, and activity diary data.

This thesis is concerned with the exploration of event-based data, and in particular the identification and analysis of sequences within them. Sequences are interesting in this context since they enable the understanding of the evolving character of event data records over time. They can reveal trends, relationships and similarities across the data, allow for comparisons to be made within and between the records, and can also help predict forthcoming events. The presented work has researched methods for identifying and exploring such event-sequences which are based on modern visualization, interaction and data mining techniques.

An interactive visualization environment that facilitates analysis and exploration of event-based data has been designed and developed, which permits a user to freely explore different aspects of this data and visually identify interesting features and trends. Visual data mining methods have been developed within this environment, that facilitate the automatic identification and exploration of interesting sequences as patterns. The first method makes use of a sequence mining algorithm that identifies sequences of events as patterns, in an iterative fashion, according to certain user-defined constraints. The resulting patterns can then be displayed and interactively explored by the user. The second method has been inspired by web-mining algorithms and the use of graph similarity. A tree-inspired visual exploration environment has been developed that allows a user to systematically and interactively explore interesting event-sequences. Having identified interesting sequences as patterns it becomes interesting to further explore how these are incorporated across the data and classify the records based on the similarities in the way these sequences are manifested within them. In the final method developed in this work, a set of similarity metrics has been identified for characterizing event-sequences, which are then used within a clustering algorithm in order to find similarly behaving groups. The resulting clusters, as well as attributes of the clustering parameters and data records, are displayed in a set of linked views allowing the user to interactively explore relationships within these.

The research has been focused on the exploration of activity diary data for the study of individuals' time-use and has resulted in a powerful research tool facilitating understanding and thorough analysis of the complexity of everyday life.

Populärvetenskaplig sammanfattning

Utforskning av sekvenser i händelsebaserade data

Denna avhandling presenterar metoder för att studera och analysera händelsebaserade data med hjälp av modern datorgrafik och algoritmiska beräkningar.

Händelsebaserade data påträffas dagligen i många discipliner och används för olika ändamål. Data är samlingar av sekvenser som består av händelser som sker vid en viss tid och har en viss varaktighet. Exempel på händelsebaserade data är sjukjournaler som redogör för en patients sjukdomshistoria, Internetsurfningregister, biografiska redogörelser, redogörelser för förflyttningar eller karriärer, samt tidsanvändningsdata i form av aktivitetsdagböcker som är register över hur individer använder sin tid för att genomföra sina dagliga aktiviteter, vilket är den typ av data som står i fokus i detta forskningsarbete.

När man studerar händelsebaserade data i allmänhet, och tidsanvändningsdata i synnerhet, är det av intresse att identifiera sekvenser av händelser, eller aktiviteter, som sammantaget uppvisar ett specifikt beteende. Detta kan, till exempel, vara sekvenser som är gemensamma för många och ofta förekommer på ett liknande sätt, eller som är unika för ett fåtal och avslöjar avvikande mönster. Genom att identifiera och synligöra sådana sekvenser blir det möjligt att hitta samband och trender samt genomföra jämförelser inom och emellan dataregister. Inom tidsanvändningsstudier handlar detta om att studera hur individer bygger upp sina dagar, hur de arrangerar sina dagliga projekt, kombinerar alla sina måsten och pusslar ihop sina vardagsliv. Det vanliga sättet att analysera tidsanvändningsuppgifter är att skapa rapporter av sammanfattande statistik, i form av tabeller och diagram, över total tid som tillbringas på olika aktiviteter. Även om denna metod ger värdefull övergripande information, försummar den också viktiga egenskaper som gömmer sig i tillgängliga data. Detaljer för när, hur många gånger, hur länge och i vilken ordning de olika aktiviteterna genomförs förblir dolda.

Forskningen som presenteras i denna avhandling har fokuserat på att utveckla metoder för att visuellt analysera händelsebaserade data, särskilt aktivitetsdagböcker, som synliggör och utnyttjar deras inneboende sekventiella karaktär. Representationer som är skräddarsydda för den särskilda datatypen har konstruerats. Dessa i kombination med grafiska gränssnitt, interaktions- och filtreringstekniker ger en användare möjligheten att fritt utforska och studera data strukturen. Utöver detta har olika algoritmer för datautvinning varit föremål för forskning i syfte att automatisk kunna identifiera intressanta sekvenser inom samt genomföra jämförelser och gruppera dataregister med hänsyn till de identifierade sekvensernas likheter. Alla utvecklade metoder har kombinerats med visualiserings- och interaktionstekniker för att effektivt presentera och tillåta interaktiv utforskning av data och resultat. Arbetet har resulterat i ett kraftfullt forskningsverktyg som möjliggör meningsfull och ingående analys av händelsebaserade data.

Acknowledgements

First and foremost I would like to thank my supervisors Matthew Cooper, Kajsa Ellegård and Anders Ynnerman for their guidance, advice, support, inspirational and motivating discussions, and all the proof readings throughout these years. I couldn't have been here without your help.

I would also like to thank all my friends and colleagues at NVIS and VITA for all the discussions and suggestions during this time. A special thanks goes to Karljohan Lundin Palmerius for his additional help with LaTeX, to Jimmy Johansson for our collaboration and fruitful discussions and to Camilla Forsell for being my collaborator, my friend and for always being willing to listen.

Many thanks are also directed to Eva Skärblom for helping and arranging all the small and not so small things that are often taken for granted.

Last but not least I would like to thank my friends and my entire "outstretched" family who, whether near or far, have always supported me and shown interest in my work.

***

This work has been supported by The Swedish Research Council (Vetenskapsrådet), the Swedish Energy Agency (Energimyndigheten) and Brains & Bricks.

Contents

Complete list of publications

Kajsa Ellegård and Katerina Vrotsou. Capturing patterns of everyday life - presentation of the visualization method VISUAL-TimePAcTS. *IATUR - XXVIII Annual Conference*, Copenhagen, Denmark, August 2006.

Katerina Vrotsou, Kajsa Ellegård, and Matthew Cooper. Everyday life discoveries: Mining and visualizing activity patterns in social science diary data. *Proceedings of the 11th International Conference on Information Visualization*, pages 130-138, Zürich, Switzerland, July 2007.

Katerina Vrotsou, Anders Ynnerman, and Matthew Cooper. Seeing beyond statistics: Visual exploration of productivity on a construction site. *Proceedings of International Conference on Visualisation*, pages 37-42, London, UK, July 8-11 2008. IEEE Computer Society.

Katerina Vrotsou, Jimmy Johansson, and Matthew Cooper. ActiviTree: Interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):945-952, 2009.

Katerina Vrotsou, Camilla Forsell, and Matthew Cooper. 2D and 3D representations for feature recognition in time geographical diary data. *Information Visualization*, doi: 10.1057/ivs.2009.30, 2009.

Katerina Vrotsou, Kajsa Ellegård, and Matthew Cooper. Exploring time diaries using semi-automated activity pattern extraction. *electronic International Journal of Time Use Research*, 6(1):1-25, 2009.

Kajsa Ellegård, Katerina Vrotsou, and Joakim Widén. VISUAL-TimePAcTS/energy use - a software application for visualizing energy use from activities performed. *Proceedings of 3rd International Scientific Conference on "Energy Systems with IT"*, Älvsjö, Sweden, 16-17 March 2010.

Katerina Vrotsou, Anders Ynnerman, and Matthew Cooper. Behaviour-driven clustering based on event-sequence similarity metrics. Submitted to *Information Visualization*, 2010.

Part A

Context of the work

Chapter 1

Introduction

We are living in the age of *too much* data. Technology is advancing day by day, with new data sources being discovered and new methods and devices that facilitate the collection of data being invented and old ones refined. The capacity for storing all this new data is continuously getting larger and more efficient which in turn means that even more data is collected. We are able to gather so much and such complex data from sources that would have been unimaginable only a few years ago. There is one thing, however, that is not improving as rapidly and that is the capacity of our human brain to process all this data that we are exposed to and extract meaningful information from it in order to reach some kind of insight. The amount of data becomes so large that it is difficult to interpret and convey. Therefore, and since we do not have the capacity to process the raw data ourselves, we are instead becoming experts in developing methods that can aid us in this task. Visualization and data mining are techniques meant to do just that, and, moreover, combinations of the two can often render even better results in terms of understanding.

Visualization is concerned with understanding and extraction of information from data. The term is not equivalent to visual representation, which is a common misconception, but visual representations are often used in order to achieve it. Visualization is the internal process that the mind undergoes in order to create a mental image of the data and gain insight into the information they contain [80]. Visualization is usually divided into information and scientific visualization. The difference between the two is mainly the type of data they are concerned with, even though they overlap to a great extent. Scientific visualization is usually associated with physical data, like medical, biological or meteorological for example, and 'real world' representations tend to be used. Information visualization can be concerned with any kind of data, such as economic, transaction, or statistical amongst others, and is not restricted to any tradition of representations, they are often entirely abstract.

Data mining is, as the term implies, to mine, to look for the 'gold' in vast amounts of data. It is the process of extracting relationships and useful knowledge from large datasets which would otherwise remain hidden. As the availability of data and data types increases so do the data mining algorithms and the types of patterns sought. There are two general types of task addressed through data mining; *predictive* and *descriptive* tasks. Predictive tasks aim at making predictions of attribute values based on other attribute values in the existing data. Descriptive tasks focus on extracting interesting patterns that characterize the properties and reveal relationships in the data [82]. Depending on the data and the task at hand different approaches apply in order to retrieve different kinds of interesting patterns. Furthermore, since what constitutes an interesting pattern is often subjective, an important component of any data mining system or algorithm is that there should be the possibility for a user to adjust settings and thus direct the focus of the search [31].

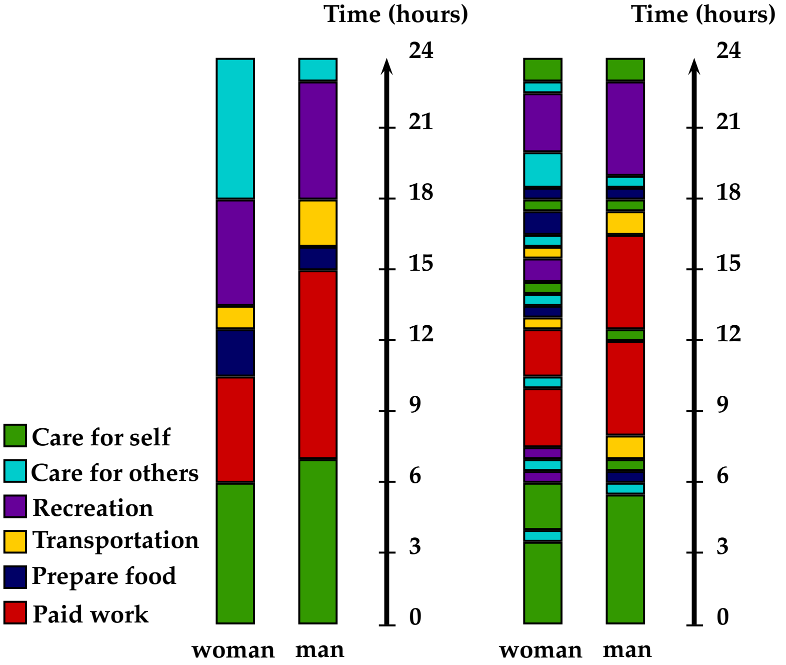
Visualization and data mining have a common goal, namely to extract and/or communicate useful information from sets of data, but they use different approaches to realise this. Visualization relies on visual representations which are good for displaying things while data mining relies on algorithms which are good for identifying/finding things. Often, the two combined can produce much more flexible and understandable results.

1.1 Everyday life studies

The underlying aim of this research has been to develop interactive approaches to understand, analyse and identify patterns in data concerned with the everyday life of individuals. Everyday life studies are often performed through the analysis of time-use survey data collected by statistics bureaus worldwide.

The motivation behind time-use surveys is that they can help map the way people use their time by considering the number of hours spent, within a given time period, on different categories of activity, for example work, personal care, childcare and housework. Time-use surveys produce data that have valuable applications in a wide variety of fields. Examples include sociology, which is also the focus of the presented research, where data of this type enable the study of such things as gender and/or age differences with respect to working hours, household division of labour, socializing and parenting among other things. In economics where working hour patterns can be of interest. In urban planning and human geography where transportation means and time spent on transportation can be studied in relation to working and/or store opening hours, for example. In policy making, in general, time-use data can aid in making better justified decisions with respect to how these may affect individuals' daily lives. It can also be useful in more simplistic descriptions such as in journalism, to account for time-use related statistics. Overall, time-use survey data can be interesting to anyone who wants to explore and compare time-use patterns of and between populations and subgroups of such.

The data collected in time-use surveys are in the form of activity diary data. Individual volunteers log all the activities they perform over a period of time, usually daily, in handwritten diaries. These diaries can either be composed of predefined time slots to be filled in, or individuals can fill in start and end times for each activity and hence provide higher time resolution in the data. Apart from the type of performed activity, information about where and together with whom each activity is performed can also be collected. The activity descriptions, places and companionships included in the diaries are then translated into a diary database with respect to some predefined coding scheme in order to enable the comparison of the diaries in a uniform language. The diaries then take the form of a set of daily sequences of activities per individual, in which each activity has a



(a) Summary representation (b) Real time-use representation

Figure 1.1: Two example representations of the daily time-use of a man and a woman. (a) Traditional summary representation that shows the total amount of time spent on different activity categories during the day. (b) 'Real time-use' representation that reveals the sequencing, timing and duration of activities during the day.

corresponding start time and duration as well as additional attribute information. Finally, demographic information about each volunteer is also usually collected, such information includes sex, marital status, number of children, education, occupation, household and individual income amongst others.

The common way to analyse time-use data is to create reports of summarizing statistics in terms of tables and charts. Usually these present the total amount of time spent on the different activity categories. Detailed graphs are constructed for various age and sex groups, or depending on marital status and employment status etc. so that comparisons between groups of different characteristics can be made. An example of a typical representation used is seen in figure 1.1(a) where the total time spent on activities of different types during a single day (24 hours) by a man and a woman is represented.

The problem with using such summarizing representations is that valuable knowledge that is actually contained in the data remains unused. Traditional representations account only for the total time spent on activities, while information such as when activities occur, how many times and for how long is not considered even though it is available. The structure and rhythm of the individuals' daily lives which is reflected in the way they interwove their activities during the day is, therefore, lost. Moreover, such representations can sometimes render misleading results as the total time allocated for certain activities may be similar for two compared groups, while the way in which this time is allocated with respect to timing, duration and repetition may be very different. Considering, furthermore, the fact that time-use data can be used as a basis for policy decisions makes this an issue that cannot be taken lightly as it would provide insufficient grounds for the task. Finally, taking into account the high cost involved in performing time-use surveys does not justify such information loss. There is, thus, a need for alternative, more efficient representations that make better use of this collected material.

An alternative approach to the representation of the activity diary data, collected through time-use surveys, is to embrace their sequential character instead of disregarding it. This can be done through considering and handling the diaries as the ordered sequences of activities that they in fact are. Representations that reveal attributes such as time of occurrence and duration, as in the example seen in figure 1.1(b), instead of hiding them are then to be preferred. Considering the data in this manner makes it obvious that diary data is a special type of sequence data referred to as event-based data.

1.2 Event-based data

Event-based data are defined as data composed of sequences of ordered events, an *event-sequence*. Each *event*, or *element*, of an event-sequence has a start time and a duration and each begins when the previous one ends. The types of event present in such a sequence all belong to a set of predefined event types, an event alphabet. An event-based dataset, D, then consists of a set of event-sequences:

$$D = \{S_1, S_2, ..., S_m\}$$

where m is the total number of event-sequences in the dataset. Each sequence, S_i for $i = 1, 2, ..., m$, is composed of an ordered list of events each of which has a start time, t, and a duration, dur. Considering the total set of event types E, each event is a triple (e, t, dur), where $e \in E$ is the event type, $t \in \Re$ is the start time in minutes and $dur \in \Re$ the duration of the event in minutes. Each event-sequence is then described by a sequence of such events:

$$S = \langle (e_1, t_1, dur_1), (e_2, t_2, dur_2), ..., (e_n, t_n, dur_n) \rangle$$

where n is the total number of events in the sequence, the *length* of the sequence, and

$$\begin{aligned} &e_i \in E && \text{for} \quad i = 1, 2, ..., n, \\ &t_i < t_{i+1} && \text{for} \quad i = 1, 2, ..., n-1, \text{ and} \\ &t_i + dur_i = t_{i+1} && \text{for} \quad i = 1, 2, ..., n-1 \end{aligned}$$

In the case of activity diary data each event-sequence in the dataset represents an individual's diary day consisting of a sequence of activities taken from a predefined activity coding scheme. The total duration of each diary corresponds to the predefined duration of the observation period and should usually be the same length for all participants in the survey. Diaries collected through time-use surveys for everyday life studies usually include records per 24 hours of a day (1440 minutes), so an additional restriction applies:

$$\sum_{i=1}^{n} dur_i = 1440.$$

Apart from activity diaries, event-based data is found in a wide range of application areas. Internet surfing records can be seen as a sequence of events (the web site visits), for example, the same can be said of working careers, historical events, personal and travel histories, medical records, work sampling records, purchase transactions, industrial process and system control records, records of attended university courses, among many others. An important feature in this definition of event-based data is the continuity of the event occurrences which should be preserved, so when handling data where short events occur with long periods of inactivity in between an idle/empty event can be used for describing these periods. Furthermore, if the time and duration constraints are relaxed even other data with an explicit ordering can be seen as event-sequences such as protein and DNA sequence data. In all cases a common factor of interest when analysing event-based data is the comparison of event-sequence records and the identification of interesting features as patterns within them.

Comparing event-sequences in a dataset reveals general trends, relationships and similarities between records and can lead to an initial classification of the data depending on these. Visualization as well as data mining techniques can be used to achieve this through, for example, representations, highlighting and sorting approaches, or automatic comparison and classification methods. Inspecting an activity diary data representation created in a meaningful way can, for example, lead to observations like ‘older people tend to go to bed earlier’, or ‘women tend to spend more time doing housework compared to men’, or ‘most recreation activities occur in the evening despite age’ etc.

When further detail in the analysis is wanted, characterizing patterns can be sought in the data and their distribution studied. A pattern in this context is usually defined as an identified sub-sequence of events that expresses some interesting attributes, which may be frequency or infrequency of occurrence, repetition, or other distribution characteristics.

Such patterns can also be identified visually through the creation of appropriate representations and the use of filtering and interaction techniques, or discovered through the use of automatic pattern identification methods, or even by combining the two into interactive visual data mining approaches. In activity diary data, for example, the sub-sequence 'wake up', 'eat breakfast' and 'read the newspaper' could be identified as a pattern due to its frequency of occurrence, studying this pattern's distribution across the data could then reveal further sex and age differences. Similarly in process control data a sequence of events leading to a system failure could be identified as a pattern due to its infrequency of occurrence and further study of its distribution might help in avoiding such failures in the future.

1.3 Research challenges

The temporal nature and multidimensionality of event-based data is what makes their study both complex and interesting. The focus of the work presented here is activity diaries but the research approaches aim at a wider applicability. The overall goal and major challenge of this thesis is the creation of interactive approaches that combine visual and automatic methods in order to reveal interesting attributes, relationships and patterns from event-based data and enable the comparison and classification of the event records based on these patterns.

The process of identifying and comparing such relationships and patterns is a challenge with respect to the choice of method appropriate to the task and data at hand. The creation of representations able to give both overview and detail and that enable interchange between the two without losing the overall context of the data is also demanding. Incorporating functionality in these representations that make it possible to distinguish and explore patterns of events within the data adds to the complexity. Furthermore, words like 'interesting', 'important', 'meaningful', and 'significant' are terms frequently encountered in this identification process and recur throughout this thesis. The definition of such characterizing attributes is largely subjective, since it depends on the question and objective of the data analysis as well as the analyst performing it, which makes their specification a complicated task. Another challenge which arises is the definition of similarity between event-sequences and the attempt to classify these according to the amount of similarity they exhibit. The specification of what attributes make two event-sequences more or less alike is also often a matter of the desired goal and the opinion of the user performing the classification. To summarize, the major research challenges addressed in the presented thesis are:

- the design and development of interactive visualization environments that facilitate analysis and exploration of event-based data

- the research of methods that enable the identification of interesting patterns in such data

- the specification of similarity and comparison of event-sequences based on the patterns they exhibit.

1.4 Thesis overview

This thesis consists of three parts. Part A, which includes this introductory chapter and chapter 2, introduces the research work and presents some background to it. Part B, including chapters 3, 4 and 5, summarizes the results and contributions of the presented research work. Finally, part C consists of the research papers which are included in this thesis.

Chapter 2 introduces the research areas that have inspired the presented thesis work. First, the time-geographical framework is presented, and its applicability to the conceptualization of event-based data is reviewed. Second, visual and data mining approaches to the identification of sequences and their relation to event-based data are reviewed. Finally, clustering, is discussed, as a technique to classify event-based data based on event-sequence similarity.

Chapter 3 presents the visual analysis tool, called VISUAL-TimePAcTS, that has been successively created during the course of this research for the exploration and analysis of activity diary data. The chapter reviews the tool's basic functionality and representations and also presents how energy use information can be retrieved from the explored activity diaries.

Chapter 4 summarizes the work presented in the included research papers composing this thesis. This work in this chapter is divided into four research themes concerning the exploration of sequences in event-based data. These themes are visual analysis of event-sequences, sequence mining for pattern identification, interactive sequence identification and classification based on sequence similarity. The goals and contributions of each theme and the corresponding papers are discussed.

Chapter 5 provides general conclusions and contributions of this research work in its entirety followed by a discussion of possible future directions.

1.5 Overview of the included papers

The research work presented in this thesis is concerned with the interactive analysis of event-based data, and in particular the identification and exploration of sequences within this data. Seven research papers have been created around this theme and are the core of this thesis. They are included in part C and will be presented in chapter 4.

Paper I introduces the visual analysis tool VISUAL-TimePAcTS and demonstrates how it can be effectively used for the study of everyday life through activity diary data.

The author of this thesis and second author of the paper has performed all the development work of VISUAL-TimePAcTS and wrote the parts of the paper concerned with this.

Paper II presents an evaluation, conducted through user-based experiments, of the effectiveness of the main representation of VISUAL-TimePAcTS in performing a representative, for the data type analysis, task. The author of this thesis is first author of this paper and was involved in the experimental design, performed all the implementation work and wrote the parts of the paper not concerned with the experimental process and statistical analysis of the results.

Paper III presents an interactive sequence mining approach for identifying interesting sequences as patterns in activity diary data. The author of the thesis is first author of the paper, has performed all the algorithmic adaptation and implementation work and written the paper.

Paper IV is an extended version of **Paper III** which includes includes a querying approach for filtering the identified patterns as well as a more detailed description of the background and analysis scenarios. The author of the thesis is first author of the paper, has performed all the design and implementation and written the paper.

Paper V presents the applicability of the sequence identification approach of **Paper III** and **IV** to the exploration of productivity on a construction site. The author of the thesis is first author of the paper, has performed all the implementation work and written the paper.

Paper VI presents an interactive visual data mining approach for the systematic exploration of activity sequences based on web mining techniques. The author of the thesis is first author of the paper, designed the method, has performed all the implementation work and written the paper.

Paper VII is concerned with the classification of activity diary data based on the similarity of identified activity sequences incorporated within them. The author of the thesis is first author of the paper, designed the method, has performed all the implementation work and written the paper.

Chapter 2

Background

This chapter will give the background of the research work presented in this thesis. Throughout this entire research work the time-geographical conceptual framework has been a constant and substantial influence. The representations and interaction features developed have been inspired by time-geographical principles and these need therefore to be explained first. Following this, different techniques have been applied for extracting relationships and interesting knowledge from the data so background information concerning these will then be accounted for. Each background area topic concludes with a section that relates this area to event-based data, which are the focus of this research, through examples of their applicability within it.

2.1 Time-geography

Time-geography has become a recurring term in the field of temporal data visualization and geographical information systems (GIS). Unfortunately, the common impression is that time-geography is simply a notation system for representing movement in space over time. This is a limited interpretation of the concept and only partly true. Even though time-geography provides a method and a notation system for representing movement in time and space its definition and impact goes much deeper than this. Time-geography is a way of describing reality and at the same time a way of relating to it.

Time-geography is a conceptual framework that was formulated around 1970 by Torsten Hägerstrand, Professor of Human Geography at Lund University, Sweden, in order to bring focus to the individual as a continuous entity, inseparable from time and space, in the population studies of that time. The motivation and main components of time-geography are the individual, time, and location. The basic assumptions taken are that the individual is indivisible, that time passing can be measured and that each individual can only be located at one place at a time. Furthermore, an individual in time-geography need not necessarily be a human being, it can refer to animals, artifacts or things shaped by nature and the concepts introduced in the framework can be used to describe any population consisting of such individuals [28].

Time-geography was initially brought about as a protest against the main scientific approach of that time, not necessarily so different from those of today, according to which scientists studying social and ecological systems (be they socio-economic, regional, transportation/migration, or consumer) have a tendency to extract observation objects from their context and study them in isolation even though these are not necessarily independent. A social system, however, by default is concerned with people so ignoring them cannot give a representative view of the situation. Concerning this, in [25] Hägerstrand wrote that it is common to *"treat a population as a mass of particles almost freely interchangeable and divisible"*. He also points out that even though we cannot focus on each and every individual when studying a population there should be an attempt for members of this population to retain their identity over time and for the focus to be put on the continuous life of an individual instead of the amassed behaviour of the population. So, the individual as a distinguishable actor within a population is the first important element of the time-geographical framework. As mentioned earlier, the term individual in time-geography extends beyond the human population, due to the scope of the presented work and the data under exploration, however, an individual in this description refers to a human being.

The other important elements of the framework are time and space which are in essence inseparable, they define and restrict the individual at the same time and they can under no circumstances be escaped. Time is passing continuously, 'now' can be regarded as merely an illusion, one can almost say there is no now since it can be seen as the constant transformation of future into past [8]. What individuals do and where they are at one point in time depends on what they did a moment earlier and affects what can be done a moment later. This does not imply constant change of activity since even doing nothing demands time and space, it does, however, imply constant change in the time direction since time never stops. Furthermore, since an individual, in their material presence, can only be at a single place at a time all activities carried out by them can be lined up into fixed event-sequences and hence be used to describe their whole life path [25, 53], or parts of it, such as a day, a week etc. Using this metaphor of a path also allows for their lives to be represented graphically (*individual path*) [28].

2.1.1 Concepts and representations

Time-geography, by considering things as having a spatio-temporal dimension and the consecutive actions of an individual as a sequence of events in space-time, provides a method and a notation system for studying life *"where it's possible to gather and represent data in a way that retains their spatial and temporal context"* [26]. In order for such a representation to function effectively in describing space-time co-existence it must fulfil four demands (as interpreted from [26], p.88):

1. What is represented has to be easily related to what actually happens in reality.

2. The representation has to have a wide application area. It should be easy to move between different aggregation levels, from a micro- to a macro-level, without losing the connection and context between them.

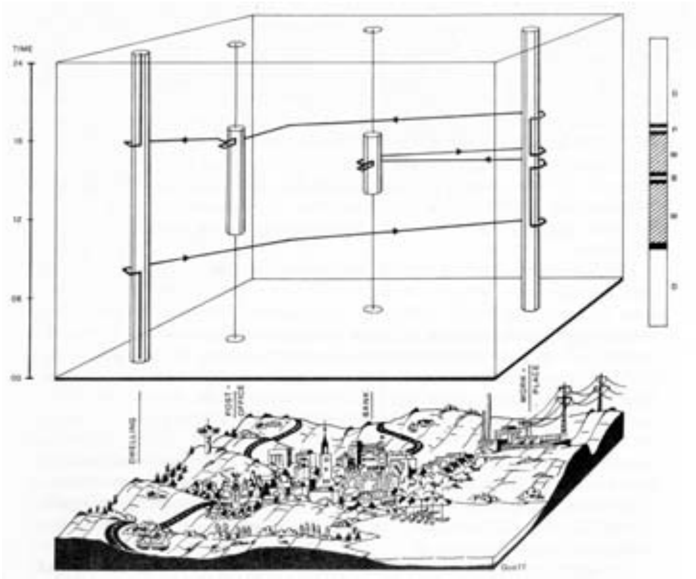


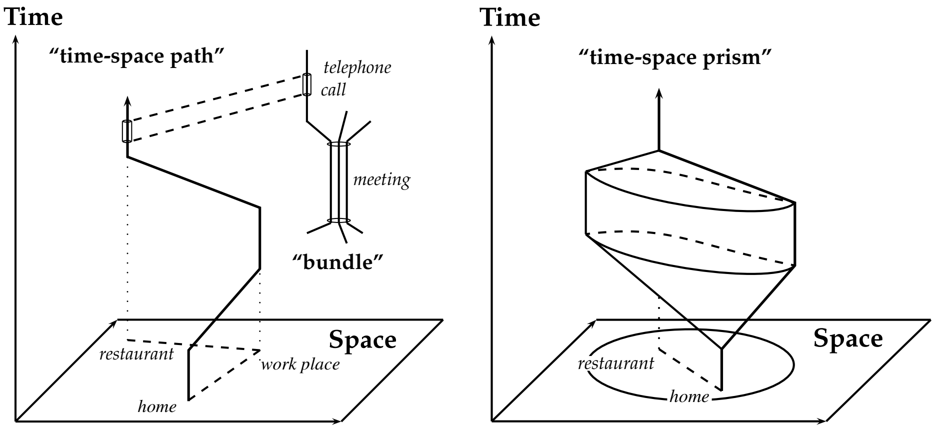
Figure 2.1: Example of a time-space cube representation. An individual's path can be drawn in a time-space coordinate system with space represented on the horizontal plane and time on the vertical axis. The movements of the individual are depicted as a continuous trajectory, the time-space path. In the example, the individual starts at home and visits his work place, a bank, his work place again, a post office, and then returns home. To the right of the figure the activities performed by the individual are displayed in a stacked bar-chart representation which does not consider their location, similar to the previously discussed "real time-use" representation (figure 1.1(b). (Source: Lenntorp, 1978 [54]. Figure reproduced with permission of Edward Arnold (Publishers) Limited)

3. The representation should be able to raise questions that would not have been asked without it.

4. The representation should lead to results whose authenticity need not be verified by observations.

In time-geography such a representation is realised through a two or three dimensional structure within which one axis represents time and the other one or two axes represent space. This coordinate system has come to be referred to as the *time-space* or *space-time cube* [42] or *aquarium* [48, 50] (figure 2.1).

The individuals daily life and movement is represented by a continuous trajectory, a *path*, following the sequence of events and is referred to in many ways: as the *individual-*, *time-space* or *space-time path*; or even as a *life*, *year*, *week* or *day path* depending on the time span it covers (figure 2.1,2.2(a)). The direction of the path is always positive along



(a) Individual paths and bundles.

(b) Individual prism.

Figure 2.2: Examples of time-geographical representations, as described by Hägerstrand [25] and Lenntorp [53]. (a) Time-space paths and bundles. To the left of the figure an individual's time-space path is displayed starting from home going to work and then to a restaurant. To the left several individuals' paths meet forming a bundle of paths in time and space. The telephone call between two of the displayed individuals is an example of a bundle in time but not in space. (b) The time-space prism of an individual displaying the time-space volume that an individual can reach at any given moment. It's size depends on the time available before the next planned event and the speed of travel. In this example the prism represents the time-space volume the individual can reach between leaving from home and before having to be at the restaurant.

the time axis and moves between locations along the space coordinates. Depending on the variable of interest or the time span considered, the path can also occupy only one dimension of space. If the individual resides at a single location, for example, the path is a vertical line as in figure 2.3.

Several individual paths, representing individuals in a *population*, can be drawn within the same time-space cube revealing a pattern of how these individuals meet, relate and interact with each other. Individuals can meet in space for certain periods of time forming groups or *bundles* of paths (figure 2.2(a)). Bundles can also be formed by meetings in time only, through telephone calls for example.

How far individuals can reach, how they move and their opportunities to participate in bundles, however, are limited by time and space as such and through other constraints. The time-space volume within the reach of an individual at any given moment forms a *prism* or *potential path space* [25, 53]. The size of this depends on the time available before having to return to a location of rest or before the next planned event, as well as the speed of travel (figure 2.2(b)).

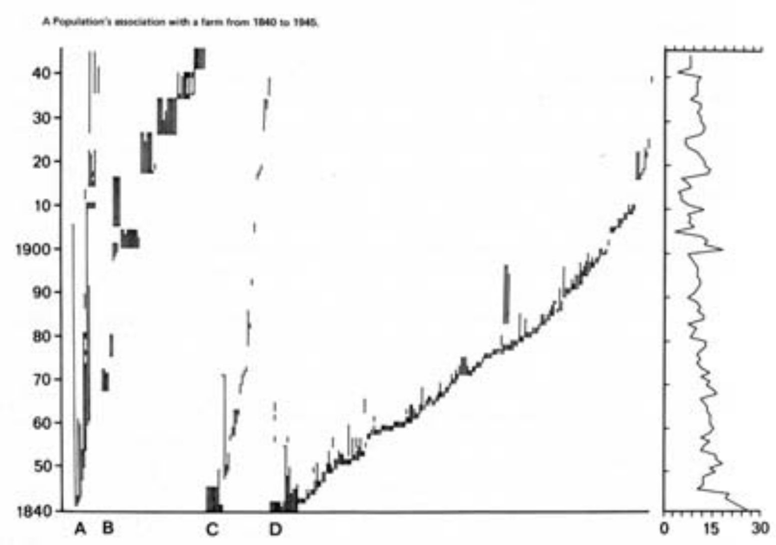


Figure 2.3: Example of a time-geographical representation. Life paths are drawn, in two dimensional space, representing a population's association with a farm between 1840 and 1945. Every vertical line represents an individual's life line while they stayed in the farm. Group A includes owners, B tenants, C lodging persons and D farmhands and maids. To the right a graph is drawn showing the number of individuals present simultaneously at the farm over time. (Source: Hägerstrand, 1978 [27]. Figure reproduced with permission of Edward Arnold (Publishers) Limited)

2.1.2 Constraints

There are three groups of constraints, which affect how individuals form their paths, identified within time-geography: 'capability constraints', 'coupling constraints' and 'authority constraints' [25].

Physiological necessities are 'capability constraints' since an individual cannot avoid the fact that they have to sleep for a number of hours at regular times and also eat at regular intervals. Constraints that restrict an individuals capacity to perform activities due to lack of resources also fall under this category. Examples of such include distance and the lack of transportation means or enough time, lack of knowledge, absence of appropriate technologies, previous experience etc. These constraints are reflected in and affect an individual's daily prism.

'Coupling constraints' apply to when, where and for how long individuals have to meet with other individuals or tools, means, services and materials. Examples are meetings of friends or colleagues, working hours, time schedules for transportation means, opening hours for different services, even internet or telephone meetings etc. In the time-geographical representations these have the form of bundles of paths.

The final group of constraints are the ‘authority constraints’ which relate to the time and space of power structures/relationships. In time-geography they take the form of so-called *control areas*, *domains* or *stations*. They are physical areas within which things and actions (events) are under the control of a single or group of individuals and are hierarchical in nature, meaning that domains can exist within other domains. Examples are the home, workplace, school, town, state, nation. Domains are represented as cylinders within which bundles can occur.

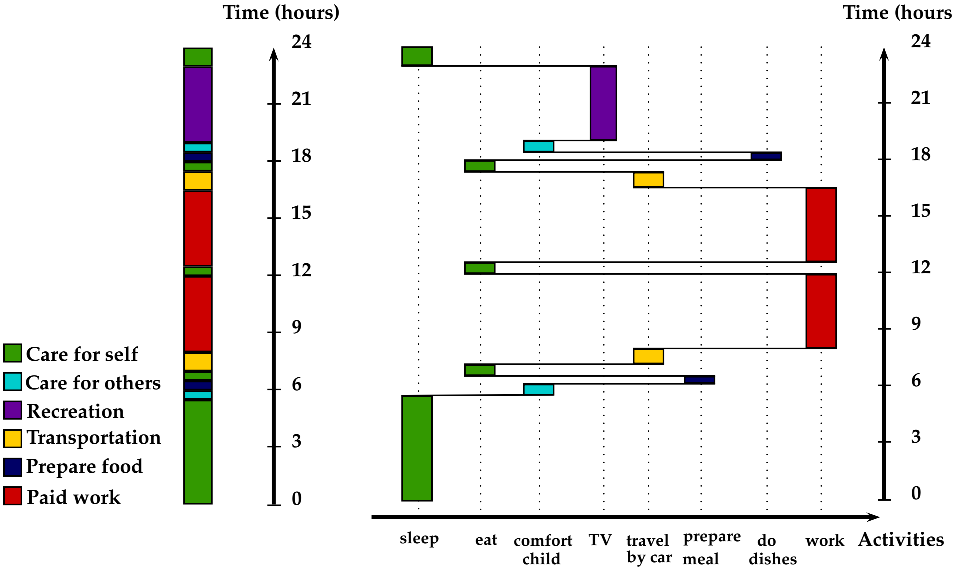
2.1.3 Time-geography and activities

From its conception the time-geographical framework has had its focus on human activity. Populations are studied by looking at how the individuals composing them co-exist in confined regions, which is realised through looking at their individual paths and how these move and intersect in space-time [28]. The actual activities of the individuals are then implicitly derived from studying these movements and interactions in space. Since individuals in a society plan their lives, days and activities, some in concert and some independent of each other, their movement in space and the bundles, of which they are part, convey information about their way of being [12, 55]. Where they go and who or what they meet as well as when these events occur reflects how they organise and live their daily life. The *where* and *when* become indicative of the *what* [70].

There are large amounts of data in the form of activity diaries available that could be used, as a complement to travel/movement data, in order to study the activity behaviour of individuals. Such activity diary data are collected worldwide by statistics bureaus through time-use surveys, as described in section 1.1. The usual approach to handle them is to calculate statistics considering percentages of time spent on different types of activity per day or per group of different characteristics, as in figure 1.1(a), for example. This approach is far from any time-geographical principles since populations are treated as a uniform mass instead of as individuals, thereby ignoring their variability and unique characteristics. Any detail and activity patterns that the data may exhibit are lost and so valuable knowledge is lost along with them.

In order to overcome the limitations of traditional statistical approaches a method for studying individuals’ everyday lives having the actual activities they perform as a starting point, instead of their geographical movements, has been introduced by Ellegård [8, 9, 10]. This method is based on the time-geographical framework treating the activities in the diaries as movements in an abstract space, from one activity to the next, allowing the use of the framework notation system and maintaining the continuity of the individual. All the concepts and constraints defined in the previous description of time-geography apply in this method also. Movement data are still available since transportation information can be included in the diaries, so the original representations can be drawn too. The difference in the data is the additional information about other activities performed during the day.

Each individual diary is a recorded sequence of performed activities and can therefore be represented by a path, a day-path or activity-path. Since the focus of the method



(a) Activity path (one dimension).

(b) Activity path (two dimensions).

Figure 2.4: Time-geographical representation of the activity path of an individual. (a) The activity path drawn in one dimension (time axis) as a vertical bar showing the sequence of activities performed by an individual during the day. Colour is used to depict the general activity type of each performed activity. The representation reveals when, for how long and how many time each activity is performed. (b) The activity path drawn in two dimensions (time and activity axis) as a vertical trajectory showing the 'movement' of an individual from one performed activity to the next. The representation reveals further detail concerning the type of activities performed by an individual and shows how activities of a certain category are broken up into more specific descriptions.

is on the activities performed, the actual geographical location need not be present in every representation so the path can be drawn in one dimension as a vertical line (or bar) (figure 2.4(a)). Details about when, for how long and how many times activities are performed by the individual are revealed using this representation. Using this one dimensional representation makes it possible for several paths to be aligned in a coordinate system and allows for visual comparisons to be made between populations. Furthermore, using the metaphor of movement of an individual from activity to activity allows for the path to also be drawn in two dimensions (figure 2.4(b))) with one axis representing time and the other an abstract 'activity space'. Representing the activities in this manner allows for further information detail with respect to the description of the performed activities to be incorporated in the path. How activities of a certain category are broken up into more specific descriptions can be shown.

The activity representation of this method is in accordance with the time-geographic principles introduced previously, in section 2.1.1. The diary representation is simple and easily related to the actual performed sequence of activities. In fact it is a direct and accurate translation of the data to a graphic representation and hence it's authenticity is indisputable. Furthermore, one can move from the detailed description of the day to the summary of time spent in different activity categories and hence look at the data at different aggregation levels which gives the viewer both overview and detail allowing for comparisons and hypotheses to be made.

2.1.4 Time-geography and GIS

The most important limitation at the time when time-geography was introduced was lack of computational capacity which had an effect on the practical use of the framework. One of the first applications showing the potential of using computers to combine geographical coordinates with demographic information in order to create location-specific data concerning populations was presented in 1955 [24]. Today with the advances in geographical information systems (GIS), which integrate cartography with database technology, as well as the increase in the amount of space-time data available, through the use of mobile phones and global positioning systems (GPS), time-geography has had a chance to make a come-back and show its potential anew [63].

There are several reasons why GIS makes an appropriate platform for time-geographic studies. The spatial organization of data and wide availability of cartographic representations within them are the obvious ones. Apart from these, other advantages, as summarized by Kwan in [51], are that data from different sources and formats can be integrated within such systems, the complexity of this data is retained while the manner in which they are represented allows a user to process them visually. Furthermore, they provide an interactive environment allowing a user to make changes in variables and view the results directly. Finally, they incorporate many navigational and multimedia capabilities making it possible to view data from different directions and perspectives with no larger effort. Examples of implementations of time-geographical approaches using GIS follow in the next section.

2.1.5 Applications in event-based data

Time-geography, as defined in section 1.2, is well suited to event-based data applications, since it conceptualizes human activity and movement in terms of sequences of events that have a spatial and temporal dimension. Any type of dataset in which each data record can be described as such a sequence can be seen as a *population* [28] and hence be represented and studied under a time-geographical assumption. So the application areas are wide and can include anything from travel or activity diary data and transportation records tracked through mobile phones or GPS, to personal histories or biographies, internet session data, medical records and even process control data and biological sequence data, just to name a few. The focus of this thesis is in interactive visual analysis of event-based data, primarily in the form of activities for studying everyday life, so concrete application examples relating to time-geography will be confined within these fields.

Activity diaries have been widely used for studying various aspects of everyday life through time-geographical analysis, many such studies have been inspired by the activity analysis methodology discussed in section 2.1.3. Within the social sciences and occupational therapy, for example, to analyse everyday life [10, 15], or concentrating on women's daily lives, health and well-being [18, 64, 88], in [13] the use of diaries as a tool for self-reflection and rehabilitation is considered. An interactive visualization application for representing activity diaries was introduced in [11], and has been the starting point of this work.

In this thesis time-geographical representations and principles combined with powerful data mining techniques have provided the basis for an interactive visual analysis tool for the exploration and analysis of activity diaries. This work will be presented in detail in the following chapters (chapters 3 and 4).

One of the first to incorporate the time-geographical framework notation into GIS-based approaches was Kwan [47, 48]. She has since done extensive work in the study of human activity with a focus on movement in time-space using time-geographical representations within 3D GIS environments [49, 51]. A summary of this work can be found in [50]. Daily activity patterns using time-geographic methodology in GIS have also been explored by Huisman and Forer [34, 35]. Different concepts of human interaction and appropriate time-geographical representations to display these have been discussed in [95]. An approach to studying internet activity data was presented in [75], where an alternative to the traditional space-time cube, the *information cube*, is introduced.

The time-space cube for studying spatio-temporal data has also been implemented in an interactive geo-visualization environment by Kraak in [42], who has used it, for example, in an interactive version of Minard's map of Napoleon's march to Russia [41], as well as for exploring epidemiological data [44]. The benefit of time-geography in analysing health risks has also been discussed by Forer [17]. The analytical benefits of combining time-geography with geovisual analytics for studying complex spatio-temporal data are discussed by Kraak and Huisman [43].

Finally, applications of time-geographic concepts on event-based data unrelated to human activity and transportation can also be found. Turdukulov et al. [83], for example, have applied the framework's notations to the study of cloud paths, while Gatalsky et al. [21] use the space-time cube to visualize unconnected earthquake events.

Even though time-geography principles have originally been developed for describing human activity and interaction, it is important to keep in mind that the concepts introduced can be applied to any dataset composed of *"indivisible and relatively durable entities that can be considered as a population"* [28]. In fact many visual representations could benefit from a time-geographic perspective due to its ability to convey complex multivariate information in a manner that is intuitive and easy to relate to.

2.2 Sequence identification

Event-based data, which are the focus of this research thesis, are a type of sequence data. *"A sequence dataset consists of sequences of ordered elements, or events, with or without a concrete notion of time"* [31]. In this context, a concrete notion of time implies an exact time-stamp indicating the initiation of each event as opposed to a relative notion of time which is implicitly retrieved from the ordering of the events. Sequence data are encountered in a large number of fields, such as shopping/transaction data, internet surfing data, process control data, biological sequences, historical-, biographical-, career- event-sequences, medical records, and, of course, activity diary data, just to mention a few.

Sequences (or sequence data) are interesting since they enable the understanding of the evolving character of records in a dataset. They can give an overall view of the regular (or irregular) behaviour of the data over time, reveal trends within the data as well as help predict future events. They allow for comparisons to be made and for the progress, over time, of events incorporated in different data records to be mapped and analysed. When analysing this type of data the patterns that are sought are, most often, sub-sequences whose distribution stands out for some reason. Sub-sequences that appear very often and/or in most of the data records may be interesting to detect, or that exhibit some sort of repetitious pattern, or even that differ from the greater part of the data. Such identification and analysis of sequences or sub-sequences finds applications in many fields.

Analysing customer transactions, in terms of the sequences of purchases they include, is a very common way to apply targeted marketing, this is especially popular in internet transactions. Items that are usually purchased in sequence are identified and when a new customer's purchases show signs of an identified pattern suggestions about the purchase of additional items are proposed by the merchant. Phrases like: *'Customers who shopped for ... also shopped for ... '* or *'Frequently bought together with ...'* are common on internet shopping sites. The same principles can also be applied to internet surfing data for directed advertising, by identifying sequence patterns of web site visits advertisements that match these patterns can be shown.

Identifying sequences of events in medical records can be extremely valuable. It is, for example, interesting to identify sequences of diagnoses that often result in a certain condition, or are common for groups of similar characteristics, or sequences of prescriptions that prove harmful or even lethal. In this way medical incidents that appear unrelated may prove to be associated, conditions may be diagnosed earlier on if there are indications of what to look for, and possibly future problems prevented. The same reasoning also applies to process control data where it is interesting to find sequences of events that often lead to a system failure and thus prevent it.

Sequences are also of interest when using activity diaries to study individuals' everyday lives. In this case one wants to see how activities are distributed across a population and how activity projects are incorporated in peoples lives; how they appear in different age or gender groups, for example. *Projects* or *activity projects*, in this context, are defined as sequences of activities that together aim at achieving a larger task or goal. They can be short or long term and of different size, time span and significance. Some projects are so routine that they are taken for granted and are hardly noticeable. Having a meal can be seen as such a project, for instance, the distinct activities one has to perform for achieving this project can include buying groceries, preparing the meal, eating it and taking care of the dishes afterwards. Projects can also be of larger scope and duration, like organizing a family vacation or getting a university degree. The number of activities that have to be performed in order to achieve these projects is larger and the activities span over a longer period of time, they can still, however, be mapped and studied.

Activity sequences, or activity projects, in daily lives are consequently sought since they are representative of how people structure their days, how they organize their to-do lists, when, how many times, for how long and together with whom they perform activities, how they piece together these activities given the individuals' available resources and existing constraints, and in these lines how they in fact live their lives. Furthermore, the same concept of activity sequences and projects applies also to the analysis of personal histories, biographies, and career paths and can aid in the understanding and explanation of peoples development and, in retrospect, social change.

Sequences are clearly an object of interest for many disciplines and for numerous reasons and these are identified using various methods and techniques. The presented work has approached the identification of sequences using visualization and data mining techniques. Visual representations of sequence data can be used to enable visual identification of patterns and trends within them. Data mining algorithms are able to facilitate the automatic identification of patterns and relationships across the data. Also the two approaches can be combined through identifying patterns automatically and communicating the results visually, or even through alternating between representations and automatic processes in order to guide the search and refine the results. The following sections will briefly describe how sequences can be, and usually are, handled in these disciplines of visualization and data mining and give some examples of their applicability to event-based data. They will provide the background upon which the presented research has been built.

2.2.1 Visual identification

Within the visualization field visual representations that display the sequential nature of the data along with other data attributes can be created in order to enable identification of sequential patterns. Interaction techniques can be incorporated within the representations to allow a user to freely explore the data; to choose between different views, sort and filter the data, and highlight attributes within them in order to get a better understanding and extract the sought results.

As with the description of the time-geographical framework, section 2.1.1, in order to create representations that function effectively they have to be intuitive to understand and easy to relate to. Also, it is important to be able to look at them at different levels of aggregation, both overview and detail, without losing the context between the two. This is along the lines of the, so called, visual information seeking mantra *'overview first, zoom and filter, then details-on-demand'*, described by Shneiderman [78], which are valuable guidelines to follow when designing visualization tools.

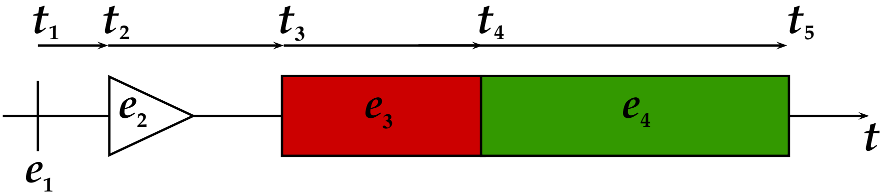


Figure 2.5: Example of graphical representations of event-sequences. The different events can be represented by glyphs of different types, sizes, and colours drawn on a time line. The duration of each event can be either implicitly derived by the time span between neighbouring glyphs, representing events, (e_1 followed by e_2) or can be incorporated in the actual glyphs (e_3 and e_4).

A sequence is often depicted as a series of events along a timeline (figure 2.5). The events are often represented by glyphs, of various types, representing different types of information. They can be glyphs of the same size and form for showing the initiation of an event (figure 2.5 e_1 or e_2). The glyphs can have different forms for denoting different types of event (figure 2.5 e_1 followed by e_2). If an event has a duration this can be implied to be the time span before the next event occurs on the timeline (figure 2.5 e_1 followed by e_2) or the duration can be incorporated in the glyph (figure 2.5 e_3, e_4). Colour can also often be used for representing different types of event (figure 2.5 e_3, e_4). However, using various shapes or colours to distinguish between types of event assumes a finite and rather small number of distinct types, since using too many can make it difficult to distinguish any belonging or similarity between sequence records.

Several sequences represented in this way can be drawn within a single display facilitating comparison between them. Sequences can be aligned by time or by a given common event making it easier to spot similarities or deviations among them. Getting an initial overview of the data can direct a user toward different hypotheses concerning the underlying patterns. Highlighting a specified sub-sequence of events in the representation based on these hypotheses then allows for focused exploration of the sub-sequence and gives further insight into whether it constitutes an interesting pattern.

Visualization techniques alone, however, are sometimes not enough. When the size of the data becomes too large, representations get cluttered and difficult to interpret. Even though sequence patterns are present within such representations they can be impossible to identify by solely visual means. Filtering techniques can aid in this but then the user has to know what to look for and what can be filtered out, this hinders free exploration and hence unexpected patterns become harder to identify. When the complexity and size of the data becomes overwhelming for the human eye it is common to adopt automatic methods to reduce the data and/or extract relationships from them.

2.2.2 Sequence mining

Sequence mining is a term characterizing a whole field within data mining concerned with the automatic identification of frequently occurring sub-sequences as patterns from large sequence datasets. Even though frequency is not always the most interesting attribute when studying sequences it is often the one used in automatic searches due to the fact that it can easily be measured.

Several concepts are needed in order to understand a sequence mining process. A *sequence dataset*, as mentioned earlier, is composed of a collection of sequences. A *sequence* is defined as an ordered list of elements or events each of which can have a time-stamp and a duration associated with it. The patterns sought in sequence mining are called *sequential patterns* or *frequent sequences.* A sequential pattern is a frequently occurring sub-sequence within a sequence dataset. A sequence b is a *sub-sequence* of a sequence a if each element in b is a subset of an element in a, if b is *contained* in a. A sub-sequence constitutes a sequential pattern (or frequent sequence) if it is satisfying a *minimum support threshold* which is a minimum frequency constraint set by a user [31, 82]. So, when performing sequence mining, one wants to discover all sub-sequences that satisfy the set minimum support threshold, meaning all the existing sequential patterns.

The trivial way to identify such frequent sequences would be to create all possible combinations of sub-sequences and test these against the minimum support threshold. This, however, is obviously a most inefficient method, both with respect to time and memory load, considering the huge number of possible combinations. Instead, the *Apriori principle*, as presented by Agrawal and Srikant in [3] for sales transaction databases and extended to sequence databases in [4], can be adopted to optimise the search. According to this principle *every sub-sequence of a sequential pattern is also a sequential pattern*, in other words, a sequence can be frequent only if all of its sub-sequences are also frequent. Using this observed inherent property of sequences implies that if a sub-sequence is infrequent then all sequences that contain it cannot be frequent and hence the sequence can be ignored. This can dramatically reduce the search space and make the discovery of patterns possible in realistic times.

Apart from frequency, other attributes of the sought sequences can also be constrained by the user in order to further reduce the search space. These are concerned with time and the way in which a sub-sequence, or potential sequential pattern, is identified in the dataset. These are referred to as sliding windows and time constraints and include a sliding window or maxspan, time window, mingap and maxgap constraints [81]:

- *Maxspan constraint.* A constraint can be set to restrict the maximum allowed time difference between the first and last element of the identified sub-sequence. A maximum duration allowed in order for an identified sub-sequence to constitute a sequential pattern.

- *Time window constraint.* A time window can be set to restrict the maximum time difference between any two consecutive elements of the identified sub-sequence.

- *Mingap & maxgap constraints.* A minimum and maximum time gap between two consecutive elements of an identified sub-sequence can be defined by the user.

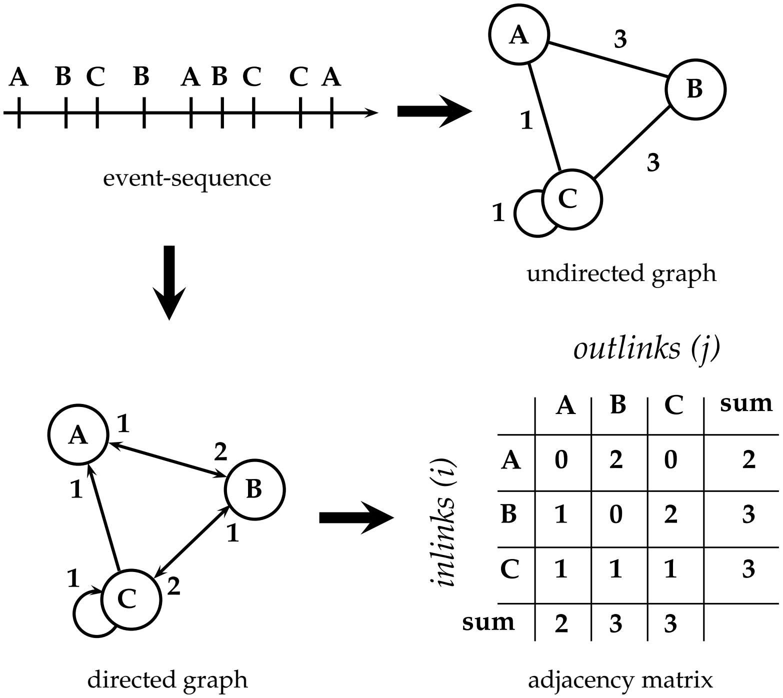
These constraints are also subject to the Apriori property in that if a sequential pattern does not satisfy the set time constraints then all of the sequences containing it will fail the support thresholds too. The Apriori property with or without time constraints forms the basis of numerous sequence mining approaches.

An alternative way to consider a collection of sequences is using graph representations. Each sequence is an entity in a dataset but the events composing the sequences are shared since they are taken from a finite 'alphabet' of events and each sequence can be seen as a transition path between a number of these events. Consequently, each distinct event, or element, within the sequences of the dataset can be regarded as a *vertex* or *node*, i, and each transition from one event to the next as an *edge*, (i, j), between the vertices. The edges could be *undirected* or *directed*, in the latter case they can also be referred to as *links* (figure 2.6). Doing this results in a somewhat simplified representation of the dataset since the concrete time notion of it is reduced, immediate ordering effects of the events are, however, preserved when using a directed graph which allows, even here, for the exploration of event-sequences as patterns. At the same time this representation also allows the identification of more complex structures within the data, such as trees, graphs and networks of events. Furthermore, using graphs allows for a more concise representation of the data, since even very large datasets consist of a finite number of nodes. The data can, therefore, be described through an adjacency matrix, by setting each entry (i, j) of the matrix equal to the number of transitions from node i to node j (figure 2.6), which in turn enables the use of matrix methods for problem solving. This can imply considerable improvement in computational efficiency.

Approaches based on the Apriori property, as previously described, can also be used within this graph representation of the data. The property, in this case, states that all sub-graphs of a frequent graph are also frequent. The patterns sought are then frequent sub-graphs, instead of sub-sequences, given a predefined minimum support threshold. Commonly, the direction between the activities is not considered using such approaches, so the identified frequent sub-graphs consist of related events that are usually found close together with no consideration of order and relationship type.

Using the notion of the graph, however, opens up the possibility of enhancing the exploration of the way in which nodes, in this case events, are linked together. The direction of the links, the relationships between the nodes and the attributes of the nodes and links are then brought into focus. The data mining area concentrating on this is referred to as *link mining* and is a combination of research in link analysis, web mining, graph mining, relational learning and inductive logic programming [22, 31] and is widely used in social networks, text and web analysis. Tasks that are common within this approach include (as described in [22]):

1. *Link-based object ranking*, the objective of which is to rank the order of objects in a graph based on its link structure.



	A	B	C	sum
A	0	2	0	2
B	1	0	2	3
C	1	1	1	3
sum	2	3	3	

Figure 2.6: Example representation of an event-sequence seen as a graph. An event-sequence can be considered as a transition path between a number of distinct events. Each distinct event within this sequence can be regarded as a node,i, and each transition from one event to the next as an edge, (i, j), between the nodes. These edges can be *undirected*, revealing only the general connectivity between the events, or *directed*, revealing detail about the transition ordering. Using such a graph representation allows for the event-sequence to be represented through an *adjacency matrix* by setting each entry (i, j) of the matrix equal to the number of transitions from node i to node j.

2. *Link-based object classification*, the objective of which is to classify objects in a graph not only based on attributes that describe the objects themselves but also their relations with objects linking to them and the attributes of these related objects.

3. *Object clustering*, aims at detecting groups of graph nodes with common characteristics.

4. *Object identification*, aims at predicting the type of an object or identifying the entity in which an object belongs to based on its attributes and links.

5. *Link prediction*, refers to the prediction of whether there is a link between two objects based on their attributes and links.

6. *Sub-graph discovery*, refers to the identification of interesting or fequent sub-graphs as described previously in this section.

7. *Graph classification*, aims at categorizing a whole graph with respect to a certain concept.

8. *Generative models for graphs*, refers to models for generating graph data based on different object and link types.

When considering (event-) sequence data as a graph and trying to identify interesting sub-sequences of events within it, a link mining task of interest is *link-based object ranking*, since the identification of strongly linked objects in a graph implies the identification of significant transitions between events, and hence also gives indications about significant sequences. The widest application area of this task is in web analysis, where the internet is represented as a graph with the nodes being the web pages and the edges links from one page to another, and the most well known approaches are the *PageRank* [67] and *HITS* (Hypertext Induced Topic Search) [40] algorithms. The objective is, given a query made on a search engine, to assign some quality score to the pages matching the query, filter out the uninteresting ones and order the remaining by this score. Both algorithms are based on the assumption that the number of links into (*inlinks*) and out of (*outlinks*) a web page give information about the importance of this page. PageRank builds on the notion that the more inlinks a page has the more important it is; the higher *rank* it gets. Furthermore, if the page has inlinks from other highly ranked pages then its importance is even greater. Based on this, "a page has high rank if the sum of the ranks of its backlinks [inlinks] is high" [67], or "the rank of a page is a weighted sum of the ranks of the pages that have outlinks to it" [7]. HITS, on the other hand, assigns two scores to each web page through a *mutually reinforcing relationship* by introducing the concepts of *'hubs and authorities'*. According to these "a good *hub* is a page that points to many good authorities and a good *authority* is a page that is pointed to by many good hubs" [40]. So, an authoritative page is one that has many inlinks and a hub one that has many outlinks. The hubs and authorities approach has been generalized by Blondel et al. [5] to being a similarity measure between vertices of directed graphs where the hub and authority scores are similarity scores between a structure graph, having two vertices and a directed edge between them, and the graph of web pages relevant to a query.

Data mining algorithms, in general, search for relationships automatically based on constraints and rules set by the user. In sequence identification, in particular, most often one seeks for patterns that are common across the data and common usually implies frequent, occurring many times. It does not matter if the data are represented and handled as sequences of events or as a graph of linked events, frequency is in one way or another, involved in the task either as a minimum support threshold or as the count of in- and out-links. Many times, however, frequency is not the desired attribute, infrequent or deviating patterns may be the ones of interest. Infrequent pattern mining can be adopted for discovering the 'uncommon' patterns by using, for example, the same methods as described previously but reverting the constraints. An infrequent pattern must then appear at most a maximum number of times.

In general, methods exist or can be created for searching and identifying patterns of well defined and measurable characteristics. However, when the user does not know a priori what they are looking for, if they do not have a starting idea of the main trends of the data, or if what they are looking for is combinations of contrasting patterns then their identification through completely automated approaches becomes problematic. Furthermore, such approaches produce very large numbers of candidates and patterns that the user then has to manually filter in order to identify the ones of interest which can be an unnecessary waste of computational time as well as a waste of the user's time. For both of these reasons combinations of visualization, interaction and mining techniques are then often of preference. Using such a combined, visual data mining approach and iterating between automatic computation and interactive visualization can allow the user to view and control the results of the algorithm as well as steer the direction of the search while the process is going on.

2.2.3 Applications in event-based data

Sequence mining approaches find many applications in event-based data since, as discussed in the beginning of this section, event-based data are a special case of sequential data.

The application of sequence mining algorithms to sequence data (event-based data) has been extensively researched since the area was introduced with the Apriori approach in [4] and its refinement in [81] and has been well documented [31]. Several examples of extensions and variations have been introduced based on the Apriori principle that use different computational approaches to retrieve frequent sequences [62, 69, 68, 96]. The problem of producing too large a number of identified patterns was addressed in [19, 20] by the introduction of user-specified regular expression constraints. Research on mining episodes from event-sequences has been performed by Mannila et al. [60, 59]. In this work event-sequences are defined using a concrete notion of time, similar to the activity sequences considered in this thesis, and episodes are sought within the sequences which are partially ordered sequences of events, comparable to the concept of a 'project' introduced earlier in this chapter.

Event-based data can be represented using graphs and graph mining approaches can then be used for identifying patterns within them. Several approaches for the identification of frequent sub-graphs have been implemented [36, 45, 93, 46]. Most of these existing approaches, however, search for (unordered) sub-structures of the data and not sequences. A thorough review of the status of frequent pattern mining including sequential and structural pattern mining can be found in [30].

Visualization and interaction techniques have been used to represent, explore and enable identification of patterns within many types of event-based data. Examples include 'LifeLines' for exploring personal histories [71] and medical records [72] which has been extended in [85, 86] with interactive alignment, filtering and ranking techniques for more enhanced exploration and identification of interesting sequences. An interactive visual environment for exploring event data in the form of hotel visits was presented in [87]. PatternFinder [16] presents an interactive visual environment for the identification of event-sequences as patterns through the creation of complex queries. In [97] visualization techniques are applied to spatio-temporal activity data for the identification of daily activity trends. Web sessions are visually explored in search of sequences in SessionViewer [52]. Web data are also explored in [94] by combining mining and visualization techniques. State transition systems are represented as graphs and interactively explored in search of patterns in [84, 73, 74].

The examples presented here are only a fragment of the available ongoing research work involving event-based datada which underline the importance of their analysis in many fields. Event-based data are found in a vast number of areas and the identification of sequences within them is an interesting and complex problem that is often dependent on the type of data and the task to be performed.

2.3 Clustering

Up to now the time-geographical framework for conceptualizing and representing sequences of events as a means to study a population has been described. The importance of sequences as a whole, and as patterns, and why their identification and study are interesting has been discussed and methods used for this identification presented. Data is explored and sequences across it are identified as patterns given some interestingness measures defined by a user. The user can explore the results using various representations and look at how the distributions of patterns behave. The next question of interest is what can be done with the acquired information. What do these patterns imply? Does the distribution of the patterns indicate similarity between the data records that incorporate them? An interesting question that arises is whether records in a dataset can be grouped based on the sequential patterns they include? Can similarity in the resulting patterns be classified in order to reveal similarly behaving groups? In the case of activity diaries this becomes a question of whether individuals in a population can be grouped based on displays of similar behaviour. As discussed in section 2.2, activity projects, defined as sub-sequences of activities that aim at achieving a larger task, are representative of how individuals structure their days and, hence, are indicative of their daily behaviour. Studying how these projects are distributed across a population and grouping the population based on the manner in which they perform their projects is therefore meaningful in the search of groups of similar behaviour. Clustering is a data mining technique focusing on tasks of this type.

Clustering is the process of dividing data into groups based on attributes of the objects in the data and their relationships. The objective of clustering is to retrieve groups, or clusters, of data objects such that the objects within a cluster are as similar, ‘close’, as possible to each other while they are very dissimilar, ‘remote’, from the objects in other clusters. Distance measures between objects are usually used to describe this similarity.

The most common types of clustering are:

- *partitioning*, where the data is divided into a predefined number of clusters based on the distance between data objects,
- *hierarchical*, which creates a successive hierarchical decomposition of the data and can be agglomerative or divisive,
- *density-based*, where data is partitioned based on the density (number of objects) in a neighbourhood.

Other types include grid-based, model-based, constraint-based and methods for clustering high-dimensional data [31]. Since the basic clustering theory is well documented and not the focus of this research work I refer the interested reader to the data mining literature, for example [31, 82], for details. The partitioning method for clustering, as described in the cited literature, will be briefly addressed here since it is relevant for the presented thesis work.

Partitioning methods for clustering aim at classifying a dataset of objects into a predefined number groups, or clusters, where each group includes at least one object and each object belongs to a single group. There are also methods where objects can belong to more than one group, fuzzy clustering, these are however outside the scope of this description. An initial partitioning of the data objects is created, either at random or using some assignment method, and an iterative technique is then used to improve the partitioning by relocating objects between clusters depending on their proximity to each other. Two typical examples of such partitioning clustering algorithms are k-means and k-medoids [31]. The objective, as mentioned before, is for objects within a cluster to be as similar as possible and as dissimilar as possible from objects in other clusters.

There are several measures to calculate the distances which define the retrieved groups, and the choice of the distance measures depends largely on the type of variables composing the data [82, 31]. Clustering variables defined on an interval-scale such as temperature (in Celsius or Fahrenheit) and calendar dates, or variables defined on a linear ratio-scale such as counts, age, weight, and height commonly involves the use of the Euclidean or Manhattan distance for measuring the proximity between objects. The same distance measures are also frequently used for variables of ordinal scale such as grades, proficiency level, and street numbers. Variables of this type have values belonging to a set of ordered states. When measuring the computer or language skills of a person, for example, these may be classed as beginner, intermediate, or advanced. These states can be replaced by a rank number reflecting their order which makes it possible to use measures such as Euclidean distance when clustering them. For nominal-scaled variables (categorical data), such as sex, and hair colour, the ratio of mismatches between compared objects is often used as a distance measure. Clustering ratio-scaled variables which do not follow a linear scale, such as the growth of bacterial populations, the common approach is to either use logarithmic transformations to even out the differences between the values or treat the values as ordinal data and replace them by their rank number. Following this, distance measures used for interval data such as the Euclidean distance can be applied. Having a dataset of mixed variables demands that the distances are computed using appropriate measures for each variable, normalizing them to the same scale, and then combining these into a single distance measure. Finally, when dealing with more complex objects that do not have a measured distance, such as in text mining, for example, where each document is represented by a vector of terms included in it, an angular distance metric such as the cosine distance is commonly used.

In the case of event-based data, where sequences can be seen as strings, meaning they are composed of symbols belonging to a predefined alphabet, such as biological sequences for example, similarity between sequences is often measured using edit distances. Similar to the mismatch distance approach for categorical data, the edit distance between two sequences is the number of edit operations (insertion, deletion, substitution) needed in order to turn one sequence into the other. Each edit operation carries a cost that is added to the total similarity (or dissimilarity) measure between the sequences, the larger the value the less similar the sequences are. The process of comparing two sequences in this manner is known as sequence alignment [6]. Computing the distances between (aligning) all sequences results in a distance matrix that is then used for clustering them. Examples of distances used in sequence alignment include the Hamming [29] and the Levenshtein distance [57].

When performing partitional clustering on a dataset the choice of algorithm and distance measure to use depends on the task, the type of clusters sought, the size of the dataset as well as the type of variables in it. These are choices concerned with how to perform the clustering. A more difficult question that often rises is what to actually cluster against: what to measure. Given a dataset of objects which of the variables describing them define similarity between the objects. When clustering a dataset of cars, for example, there are a number of straightforward choices concerning which variables it is meaningful to group the objects by. Variables like price, cylinders, acceleration, horsepower are some rather obvious options. When clustering event-based data records based on sequence similarity the definition of this similarity is not as apparent as it can depend on a number of factors.

2.3.1 Similarity definition

Sequence comparison approaches often consider the total sequence similarity and classify accordingly. While this may be meaningful when comparing biological sequences or event-sequences, in general, where the duration and timing are not important factors, when dealing with event-sequences such as activity diaries, which are the focus of this research, attributes like when events occur and how long they last must be taken into consideration. Furthermore, as the size of the set of distinct event types used to describe the sequences gets larger the cost of the edit operations become less representative. What is meant by this is that as the number of types describing the event increases, the descriptions become less separated and hence cannot be interchanged at the same cost since this would affect the measure of similarity. Substituting the event ‘walking’ by ‘bicycling’ in an activity sequence ought to have a lower cost than substituting ‘walking’ by ‘watching TV’ since in the first substitution both event types concern transportation. Finally, considering that different people describe their days with different detail and the fact that finding people that perform identical activities during the day and with identical order and duration is, in practice, impossible, it makes less sense to compare individuals whole days in order to measure their similarity. Instead one can look at how parts of an event-sequence appear similar, how individuals structure parts of their days in similar manners.

Similarity between individuals' diaries can be measured based on how the individuals incorporate activity projects (Section 2.2), activity sub-sequences, into their days. However, even examining similarity in these terms is not self explanatory. How such similarity is defined depends as much on the actual sub-sequence under study, and on how this is incorporated in the data, as on the user analysing this data. Due to the complexity of the data and because no two individuals perform identical activities and in an identical way during the day, similarity judgements can become somewhat subjective.

Considering several individuals performing the same sub-sequence of activities in different manners there is no conclusive way to decide which are more similar to each other. An activity project can appear in individuals' diaries with different durations, starting times, it can be interrupted by other irrelevant activities of various numbers, types and durations, it can also appear several times in a single individual's diary. Consequently, are two people that perform the same activity project in the morning but with very different duration more similar than two people performing it with very different starting times but otherwise identically? Or how similar are people that perform the same project with interruptions of different lengths or of different types? Does the context in which a project is performed, meaning the activities surrounding a project, have an affect on similarity estimation? One could also consider the number of interruptions or the number of activities performed at each interruption, and how similar are individuals performing projects with many short interruptions to individuals performing projects with one long interruption? There is no definite answer to such questions as they depend partly on personal judgement as well as the kind of project under exploration. Exploring the morning patterns of a population, for example, by exploring a project including waking up, having breakfast and going to work, then starting time of this projects is an important classifier. While when studying TV watching patterns then the total duration, and the number and length of interruptions may be more interesting. There is therefore a need for flexibility in the available choices of similarity measures for the task under consideration.

2.3.2 Applications in event-based data

The largest application area of clustering and sequence comparison is in bioinformatics where similarity between protein sequences is sought in order to cluster them into groups of similar functionality or structure. Sequence alignment methods are approaches commonly used for measuring such sequence similarity and many variations, algorithms and packages exist [65, 66, 32]. A review of the similarity metrics used for sequence comparison of biological sequences, both for alignment and non-aligning methods, is available in [61]. There are also several examples of approaches combining sequence comparison methods and/or clustering with visualization techniques in order to explore the similarity between sequences and clusters of such [77, 79, 76, 33, 23]. Even though biological sequences are, to a great extent, different from event-based sequences with respect to timing and durations of the sequence elements, sequence comparison methods have been used for both types of data. The sequence alignment approach has also been extended to the social sciences, where it is referred to as optimal matching. Initially the approach was used for the study of historical data [1] and, following this, applications of it include career- and travel data as well as activity diary data [2, 90, 39, 38, 56]. Applications to diary data have also been combined with the creation of representative time-geographical paths [91].

A similarity measure for comparing event-sequences based on the context in which the events composing the sequences occur is presented in [58]. The context of occurrence of an event being defined as the set of event types preceding the event and occurring within a set time frame. In activity diary data this would correspond to the type of activities being performed before changing to the current activity.

A measure of similarity of event-based data records integrated in an interactive visualization environment was introduced in [92] with applications in temporal categorical data such as medical records. In this work an event sub-sequence is used as a target record and a similarity score is computed, between the target and all other records, based on the distance (time difference) as well as the number of mismatches (number of extra or missing events) between the target and matched events. The records are then sorted based on their similarity to the target and displayed using a time-line representation allowing comparison.

Finally, visualization and interactions techniques aiming at revealing patterns within event-based data, as described in Section 2.2.3, tend to represent the data in ways that enable a user to visually identify similarity and grouping among records [72, 85, 16].

2.4 Conclusions

This chapter has provided the research context of this thesis by introducing the research areas constituting the basis of this work and discussing their applicability to event-based data which are the data type in focus. The following chapters (chapter 3 and 4) will present the new approaches and tools developed in this research work with respect to this context.

Part B

Contributions

Chapter 3

VISUAL-TimePAcTS

This research is concerned with the exploration of sequences in event-based data and in particular activity diary data collected through time-use surveys. A visual analysis tool, called VISUAL-TimePAcTS, has been developed for displaying and interacting with such data, which has been successively extended during the course of the research work. The tool has acted as an experimentation platform for testing the different attempted directions and has played a central role in this thesis. This chapter will describe the initial implementation and basic functionality of VISUAL-TimePAcTS. This basic functionality has also been part of the contributions of the first part of the presented research which is concerned with the exploration of patterns using visual means. These contributions will, however, be discussed in the following chapter.

The name *VISUAL-TimePAcTS* itself is a description of the application's focus and functionality which is **VISUAL**ization of the **Time** when activities occur, the **P**laces where they occur, the actual performed **Ac**tivities, the **T**echnologies used while performing them and **S**ocial context in which they appear, meaning together with whom they are performed. Both the data and the representations in VISUAL-TimePAcTS have their basis in the time-geographic framework as described in section 2.1.

The application combines a database with a graphical user interface (GUI) and visualization and interaction techniques. It has been developed using C++, OpenGL for the graphics, wxWidgets for the GUI. The datasets are saved in a mySQL database and the mySQL C API (Application Programming Interface) is used to connect to this database.

3.1 Data conventions

VISUAL-TimePAcTS has been designed to input and display activity diary data but has, with no larger effort, been adjusted to also handle other types of data as long as they have the same event-based nature and follow some set data conventions.

This event-based nature implies that each data record is a sequence of ordered data elements. Each data element corresponds to an *event* and must have a recorded event code, a start time and a duration. There is no constraint on the minimum duration of an event. Additional information corresponding to each single event can be included but are optional. For activity diary data each event is a performed activity and the additional information includes location, companionship attributes and appliances/technologies used while performing each activity. Consequently each data record represents some larger structure, for example a data record could be a web surfing session which is comprised of a sequence of site visits (the events). For activity diary data each data record corresponds to a logged diary day and each event to a performed activity. The whole dataset is then a collection of data records, which for activity diaries corresponds to a collection of a populations' diaries.

Following the time-geographical convention that time is a continuously changing variable that constrains an individual's actions at all times (as described in section 2.1), no gaps are allowed in the data records (recorded diaries). This means that for the period considered in a record all entries must be filled with each event starting immediately after the previous one has finished. In the case of missing values in a dataset these can be handled by replacing them with events of type *unknown*.

Finally, additional meta-information relevant to each data record (diary) can be included and used in selection queries for extracting subsets of the data. Such information, in the activity diary data that have been used in this work, includes personal details per individual, such as sex, marital status, region where they live, education, also financial information per individual and per household, housing information, information about household appliances and more.

3.2 Coding scheme

A hierarchical coding scheme has been used for translating the handwritten activity diaries, as first presented in [10]. The coding scheme is composed of a set of predefined numerical codes categorized into 5 levels of detail (LOD) with respect to the description of the performed activity. At the most general LOD activities are grouped into 7 main categories: (1) *care for oneself*, (2) *care for others*, (3) *household care*, (4) *reflection/recreation*, (5) *transportation*, (6) *prepare and procure food*, (7) *gainful employment or school*. Each activity description then at level n is broken down into a more detailed description at level $n-1$. So, for example, 'meal' at LOD 3 is broken down to 'breakfast', 'lunch', 'dinner', 'supper', 'snack' at LOD 2.

The described activity coding scheme is more detailed than the ones traditionally used in time-use surveys. A translation to this coding scheme is preferred when importing diary data from these more general surveys since it enhances the exploration possibilities. Alternative coding schemes can, of course, also be imported.

3.3 Functionality

VISUAL-TimePAcTS provides functionality to interact and explore different aspects of the data. Complex queries can be created through the GUI where a user can select between various options representing different attributes of the data. After a selection has been interactively composed the database is queried and the population sub-set matching the selected criteria is retrieved and displayed. Several displays representing different information about the data have been implemented during the course of this research. The central representation of the diaries is a time-geography inspired one which displays the performed activities in the context of the individuals' daily lives. Different types of graphs showing frequency of occurrence of the various activities are also available. These are the basic representations which have been constantly present in all parts of the research work and will be described in detail in this chapter. Additional representations have been designed and implemented in combination with the developed research methods, these will, however, be discussed in the following chapter as they are best presented together with the contributions of each approach applied.

After an initial selection has been made and displayed, further filtering and highlighting of the displayed population is possible in order to reveal various characteristics of the data. Details about the data records and individual elements composing each record can also be retrieved on demand.

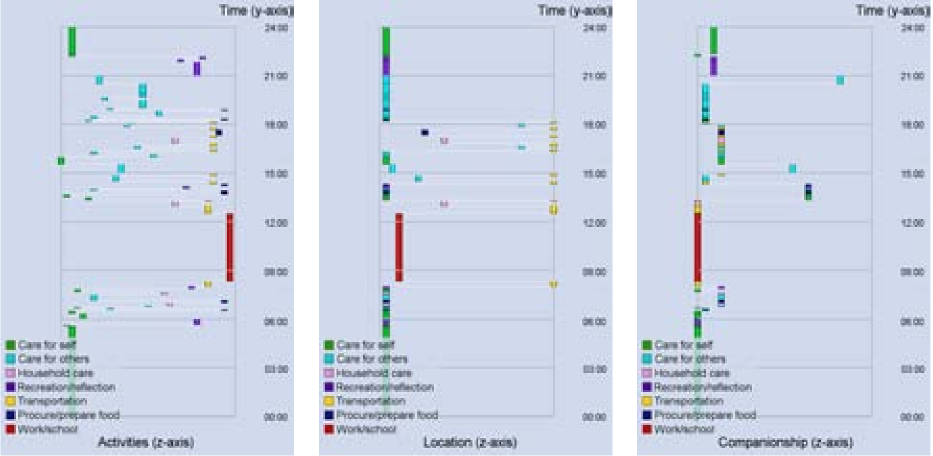
Finally, statistical information can be calculated both in terms of tables and graphs in order to give a summary of the data. Such summarizing tables can be created and saved as comma separated value files which enables their importation into other statistical software packages where they can be analysed using traditional time-use approaches. This makes VISUAL-TimePAcTS a valuable complementary tool for time-use researchers.

3.4 Visual representation

The main representation used to display and interact with activity diary data throughout this research work has its basis in the time-geographical principles, as described in section 2.1, and it will, therefore, be referred to as the *time-geographical representation*. The other basic representations initially used are line and bar graphs of different type and time resolution for displaying frequency of occurrence of activities as well as energy use information relating to the performed activities.

3.4.1 Time-geographical representation

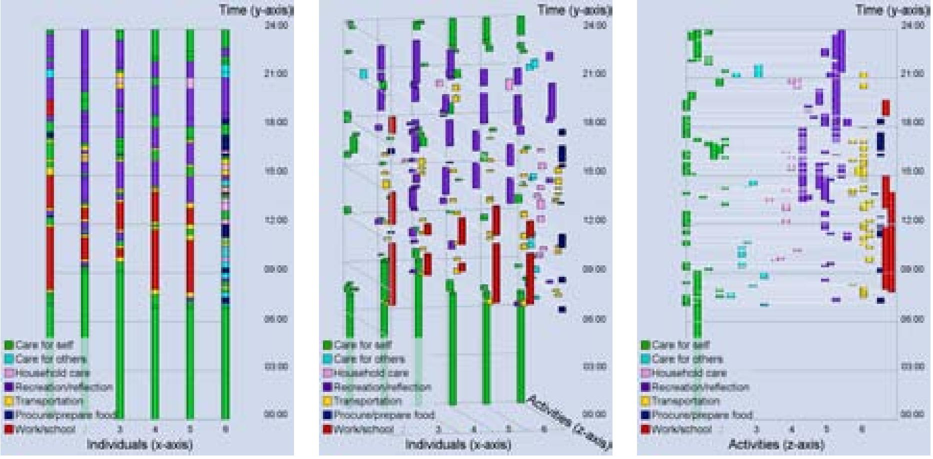
The time-geographical representation is primarily used to show *what* activities are performed and *when*, but it is also possible to show *where* and together *with whom* these activities are performed. This information is drawn using continuous trajectories within a coordinate system where the individuals are represented on the x-axis, time is represented on the y-axis and the z-axis is used to interchange between displaying activity, place or companionship (figures 3.1, 3.2). Each individual's diary is represented by a trajectory, the *individual* or *activity path*, which is the sequence of the activities performed, places visited or companionships shared by an individual during the course of the day (figure 3.1). The representation for all three features of the diary data is the same so in the remainder of this description the activity representation will be the focus but the reader should keep in mind that the same descriptions apply for place and companionship as well.



(a) Activity representation. (b) Location representation. (c) Company representation.

Figure 3.1: Representation of an individual path in VISUAL-TimePAcTS. Each individual's diary is represented by a continuous trajectory depicting the sequence of the activities performed, locations visited or companionships shared by an individual during the course of the day. Time is represented on the vertical axis and one of the three possible activity variables is represented on the horizontal axis. Each trajectory in the representation is composed of a series of tubes each of which represents a distinct activity. The position along the y-axis (time axis) and length of each tube corresponds to the starting time and duration of the performed activity. The position of the tubes along the z-axis depends on the variable represented on this axis and can correspond to performed activity type, visited location, or shared companionship respectively. (a) Activity representation revealing the interchanging between different performed activities during the day. (b) Location representation revealing the interchanging between different places where activities are performed. (c) Companionship representation revealing the interchanging between different companionships shared while performing activities.

Each trajectory (activity path) in the representation is composed of a series of tubes each of which represents a distinct activity. The position along the y-axis (time axis) and length of each tube corresponds to the starting time and duration of the performed activity. The position of the tubes along the z-axis depends on the variable represented on this axis and can correspond to performed activity type, visited location, or shared companionship respectively. The radius of the tubes in each individual's activity path is proportional to the size of the screen and the number of people present in the representation, but never exceeds $0,9\%$ of the width of the canvas. In the case of larger populations than the ones used so far overlaping effects can become a problem which will require addressing, potentially using blending techniques. The tubes, representing activities, are joined together by connecting lines in order to emphasize and conserve the connectivity of the day. Colour is used to



(a) Front 2D view (b) Rotatable 3D view (c) Side 2D view

Figure 3.2: Main time-geographical representation of VISUAL-TimePAcTS. Several individuals' trajectories are drawn beside each other in the same display allowing the exploration of whole populations' diaries and enabling comparisons between them. The individuals are evenly spread along the x-axis and can be sorted with respect to up to two of the variables sex, age, family and region. Time is represented on the y-axis and activities (or locations or companionship) on the z-axis. Colour is used to represent activity category.

signify the general category of activity and black cylinders can be turned on at the starting edges of each tube as delimiters in order to distinguish between different activities of the same category and thus same colour (figure 3.2, 3.3). The colours used for the 7 main categories are: (1) green for *care for oneself*, (2) turqoise for *care for others*, (3) pink for *household care*, (4) purple for *reflection/recreation*, (5) yellow for *transportation*, (6) dark blue for *prepare and procure food*, and (7) red for *gainful employment or school*.

Several individuals' trajectories can be drawn beside each other forming a time-activity cube, similar to the time-space cube described in section 2.1, in order to explore whole populations' diaries and make comparisons between them. The individuals are evenly spread along the x-axis occupying the total drawing span (figure 3.2 and 3.3).

The activity paths can be sorted on the x-axis with respect to up to two variables according to user preference. The sorting variables are sex, age, family and region and the default ordering is by sex and age in order to highlight gender and age differences. So, in figure 3.3, for example, women are to the right and men to the left, and each subgroup is ordered by ascending age from right to left.

The time-geographical representation seen from the front resembles a stacked bar chart giving information about when, how many times and for how long activities are performed. Which distinct activity is performed is not revealed at first glance in this view

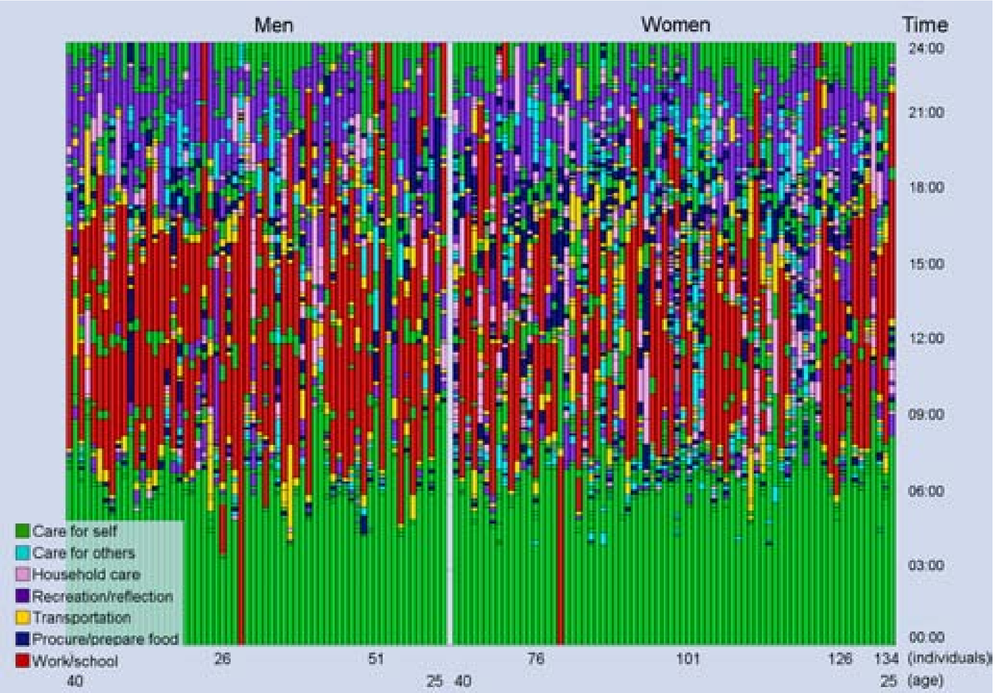
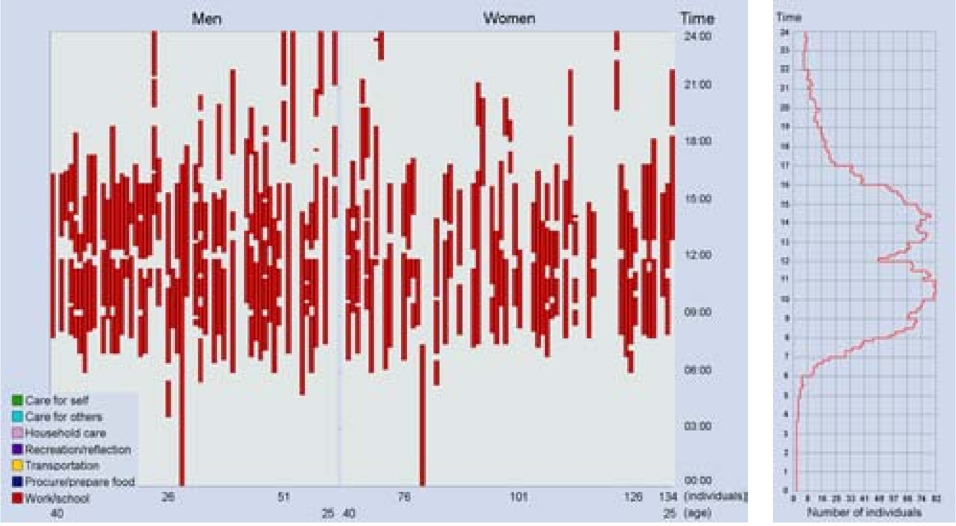


Figure 3.3: Front 2D view of the time-geographical representation of a population subset in VISUAL-TimePAcTS. The diaries of individuals aged 25 to 40 are included in this subset. Women are drawn on the right and men on the left part of the representation, both ordered by ascending age from right to left.

(figure 3.2(a)). Seen from the side the activities of each general category are broken down into more detail showing the variation of the day (figure 3.2(c)). This view, in its static form, is meaningful when studying single individuals since overlapping effects occur otherwise. The representation can also be rotated freely, enabling comparison of a limited number of individuals (figure 3.2(b)). Apart from rotation, the representation can be moved and scaled in order to study the data from different perspectives.

Details on demand can be retrieved through selecting individual activities in the representation. Information about the start time, duration and type of activity are shown in a tool-tip while more detailed information about both the activity and the individual performing the activity can be listed in the GUI.

Series of single activities or ranges of activities can be isolated and highlighted across the data in order for their distribution to be studied more closely. In figure 3.4(a), for example, the activity ‘work’ is highlighted in the time-geographical representation showing its distribution across the population in the context of the individuals’ days. Highlighting



(a) Time-geographical representation with activity *work* highlighted.

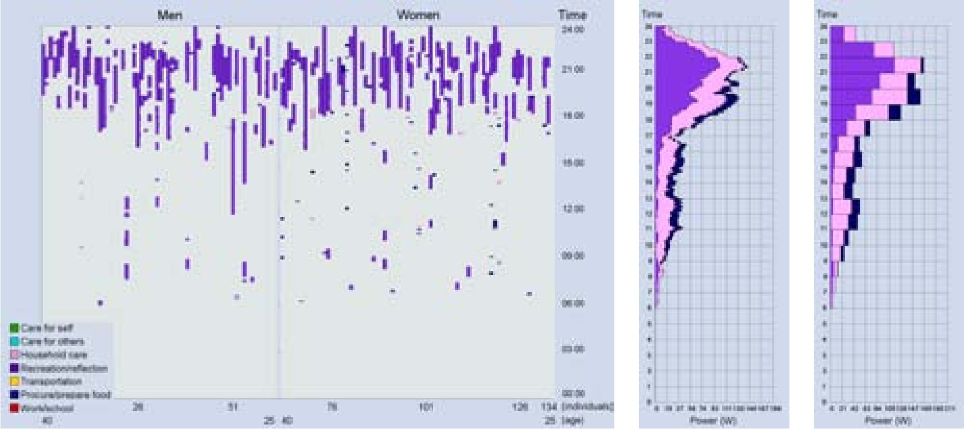
(b) Frequency graph.

Figure 3.4: Illustration of the highlighting functionality and frequency graph within VISUAL-TimePAcTS for a population subset aged 25 to 40. Highlighting selected activities in the representation reveals their distribution across the population within the context of the individuals' days while the occurrence trend of the activities is revealed through the frequency graph. (a) The activity *work* is highlighted showing the general working trend of the selected population with respect to when, how long and how many times the activity is performed. (b) The frequency graph of activity *work* is drawn showing the total number of individuals performing the activity during all hours of the day.

a set of activities in this manner can reveal where these activities appear close together and/or in sequence thus making possible the exploration of activity patterns.

3.4.2 Frequency graphs

Apart from the time-geographical representation a number of frequency graphs can be drawn within VISUAL-TimePAcTS. Single or multiple activities can be specified and graphs of their frequency of occurrence during the day drawn (see figure 3.4(b)). Graphs of different types (bar and line graph, filled or not filled graph) and resolution (5, 10, 15, 30, 60 minutes) are available for this task (examples of the various types are shown in figures 3.4(b), 3.5(b), 3.5(c)). This view makes trends of behaviour of the population more apparent and acts as a complementary visual aid in the exploration of the diaries. In figure 3.4(b), for example, the frequency graph of activity 'work' is drawn, making the performance trend curve of the activity in the population apparent. The activity starts



(a) Energy consuming activities highlighted.

(b) Load curve.

(c) Load curve.

Figure 3.5: Energy use representations within VISUAL-TimePAcTS for energy types: TV (purple), washing (pink) and dishwashing (dark blue). (a) Energy consuming activities highlighted within the context of the individuals' days in the time-geographical representation of VISUAL-TimePAcTS. (b) Mean power demand load curve computed per individual with a 5 minute time resolution and drawn as a line graph. (c) Mean power demand load curve computed per household with a 60 minute time resolution and drawn as a bar graph.

occuring around 6 o'clock and has its peak at 10, lunch breaks take place around 12 and the activity occurrences start decreasing from 15 and onward.

3.5 Energy use

The use of diary data for representing and studying daily activity also opens up the possibility for other types of analysis. The diary data include information about the type of activity performed and also the type of companionship present, while performing each activity, be that another person or an appliance. This makes possible the extraction of additional information from the same dataset namely information concerning energy use of individuals, households and whole populations.

A model for computing load profiles for household electricity and hot water use based on activity diaries was created by Widén [89] and has been partly incorporated into VISUAL-TimePAcTS [14]. For a detailed description of the model the interested reader is referred to Widén et al. [89], in this section the model will be briefly introduced and the parts implemented in VISUAL-TimePAcTS described.

According to the developed model, activity diary data are converted to energy load profiles by connecting each activity to an energy-use category and a corresponding energy-use pattern. Each such category is described by a number of parameter values corresponding

Appliance/Activity	Power (W)	Max runtime(min)
Audio	100	-
Computer	100	-
TV	200	-
Cleaning	1000	-
Ironing	1000	-
Drying	1650	90
Washing	490	130
Dishwashing	430	160
Cooking	1500	-

Table 3.1: Parameter values for electricity demand used in VISUAL-TimePAcTS. The values can be altered within the application GUI in order to explore alternative energy consumption scenarios.

to standard powers and runtime of appliances used for performing activities within it [89]. Five types of energy demand in relation to activities are identified in the model and two of them are considered in the VISUAL-TimePAcTS application:

1. Power demand not defined by activities. This concerns cold appliances' energy consumption which is independent of the individuals' performed activities and hence not incorporated in VISUAL-TimePAcTS.

2. Power demand constant during activity. This type concerns activities consuming energy while performed such as cooking, ironing, cleaning, watching TV, using audio appliances and computer, showering and cleaning. This type can be directly extracted from the activity diaries and is incorporated into VISUAL-TimePAcTS.

3. Power demand constant after activity. This type concerns energy being consumed after an activity has been performed, it refers to activities such as using the washing, dishwashing and drying appliances, and is also included in VISUAL-TimePAcTS.

4. Power demand constant during activity with time constraint. This refers to activities which consume energy for a certain time in their beginning and then no more energy is consumed even though the activity continues. An example of this type is filling the bath tub to take a bath where energy is consumed while filling the tub. This type is currently not considered in VISUAL-TimePAcTS.

5. Activities with time-dependent power demand. This refers to lighting and is similar to type 2 but the power varies with time, depending on the month of year and the time of day. Lighting information is currently not considered in VISUAL-TimePAcTS.

Based on these energy use types and their corresponding parameter values for power consumption, power demanding activities can be highlighted within VISUAL-TimePAcTS and load curves computed with different time resolutions both per individual and per household (see figure 3.5). The parameter values for electricity demand used in VISUAL-TimePAcTS are in accordance with the original model [89] and are listed in table 3.1. These can be changed by the user within the GUI in order to test alternative energy use scenarios in the represented population. Hot water, lighting and cold appliances have, so far, not been included in the model implemented in VISUAL-TimePAcTS.

Chapter 4

Exploring sequences in event-based data

The previous chapters have presented the reader with the research and application context of the thesis. This chapter will put the new approaches and the developed tools into this context and highlight the contributions accounted for in the appended papers.

The overall aim of this research is to equip the users who analyse time-use data with advanced visual data mining tools for the interactive study and identification of relationships and patterns within this type of event-based data. Patterns in event-based data are often identified as sequences or sub-sequences of events, as discussed in section 2.2. The identification of such sequences is of interest to many fields because of the information they can convey about the composition and behaviour of the data. Studying sequences can give a better understanding of how data change over time, they can reveal trends across a population as well as help predict future events based on previously identified correlations. Sequences that are frequent, that are common to many, or unique to only a few are factors that can make them interesting patterns to study. The ordering, timing and duration of events making up a sequence and the context in which events occur within larger sequences are attributes of high relevance in this identification of sequences as patterns.

The presented research has been performed using event-based data in the form of activity diaries. As discussed in the introduction (chapter 1) an important area in which the analysis of activity diaries finds applications is time-use research and the social sciences for the study of everyday life. Studying what activities individuals engage in during the day gives insight into how they structure their daily programmes and consequently how they live their lives. Making sense of this structure involves understanding how activities are combined together by individuals to form patterns of activity. Consequently, patterns in such data arise as sub-sequences of activities that, taken together, aim at accomplishing some larger task, an activity project, and exhibit an interesting distribution. Patterns of activity can also imply single activities that are performed regularly or by many at the same or similar times, for example eating or sleeping patterns across populations making the trends or groupings they form also interesting patterns to study. In both cases the common feature of interest is the way in which activities appear within the context of the individuals' days and the way in which they form activity sequences. These are, therefore, features that need to be identified in order to meaningfully analyse activity diary data.

In this research work, methods have been developed that aim to reveal such patterns and study them as well as their distribution in the context of the data at hand. These methods are described in detail in the papers included in part C of this thesis and are the core of this research. A summary of the motivation, aims, results and contributions of each of these methods will follow in this chapter, which is divided into four research themes: visual analysis of event-based data, algorithmic sequence mining, interactive sequence identification and classification based on sequence similarity.

4.1 Visual analysis of event-based data

The first part of the research work is concerned with the use of visualization and interaction techniques in order to facilitate the exploration and analysis of event-based data, and in particular activity diary data, as well as to enable the visual identification of patterns within them. An additional concern has also been to investigate the efficiency and effectiveness of the developed methods in performing a representative task within this context. These issues are addressed in papers **I** and **II**.

The core reason why a method for analysing activity diaries based on interactive visualization is important to develop is because it has been missing so far. Interactive visual data mining for the analysis of activity diaries has the potential to significantly improve the effectiveness of the research methodology, and even enable the introduction of new research paradigms, in time-use research and the study of everyday life. The current research approaches used involve statistical analysis of the data and representations of the results in terms of tables, graphs and charts of percentages and total times. While this approach is both useful and needed it demands knowledge of how such statistical analyses are performed, and also generalizes the data, often failing to reveal their full complexity. Methods which facilitate an interactive approach that allows an analyst to explore the full information span and get immediate feedback at each step of the exploration are sparse.

The use of interactive visualization approaches can complement this type of time-use analysis and lead to more well-informed results. In order for this to be possible, representations are needed that reveal the sequential character of the diaries while at the same time providing the possibility to aggregate between different levels of detail. Furthermore, such representations should make it possible to identify patterns and groupings within and between the diary records. Also, they should be able to highlight important features and aid the analyst to reach insight about their distribution. Another important issue that should be considered is the flexibility of choice as to what one finds interesting. An analyst should, hence, be able to select subsets of the data and interact with it in a way that raises hypotheses and does not constrain the exploration and discovery process.

Finally, when creating such a visual analysis system it is important to keep in mind what the objective of the analysis is, as well as what are the typical tasks performed to achieve this objective. This places demands on the system's design and should involve a thorough evaluation as to whether the requirements of the analysis are successfully addressed.

4.1.1 Aims

The aim of the first part of the research has been the development of a visual analysis software platform and end user application based on advances in visualization, and interaction techniques tailored for the time-use study of a population through activity diary data. This tool should:

- promote exploration without concealing information in the process,
- include representations of the data that both reveal their inherent characteristics but also make it possible to generalize,
- provide functionality to uncover hidden relationships and enhance the identification of patterns of activity incorporated in the data as interesting activity sequences.

Sequences, however, are not easy to spot if they are not identical for all subjects in a population. They may start at different times, have different durations and be interrupted by activities of different number and durations. Furthermore, this identification task becomes more and more demanding as the data size increases. Visual representations should, therefore, be combined with interaction and filtering techniques that can aid this identification of sequences.

Finally, there is a need to test whether the methods adopted to facilitate the analysis of activity diaries are effective in performing the tasks they are supposed and designed to perform. A representative task in the context of uncovering relationships and similarities between individuals in a population is the search for activities performed concurrently by them, since this can signify a common trend. Therefore, a second aim has also been to evaluate whether the implemented representation and interaction scheme is suitable for performing such a task.

4.1.2 Results

The basic visual representations and functionality adopted to perform analysis of activity diary data have been implemented within the interactive visual analysis tool, VISUAL-TimePAcTS, described in the previous chapter (chapter 3). The discussed research aims have been addressed in Paper **I** through building functionality into VISUAL-TimePAcTS to enhance this analysis.

Sophisticated queries can be created interactively in VISUAL-TimePAcTS allowing the analyst to freely select a sub-population of interest to focus on. A representation has been implemented that retains the full sequential character of the activity diaries and is explained in detail in the previous chapter (chapter 3). Each diary is displayed as the sequence of activities it is composed of and a population of diaries is displayed as a collection of such sequences as shown in figure 4.1. This representation combines overview and detail through showing a whole population without hiding the individual, while at the same time retaining the day context of the diaries. Colour is used for representing the general types of activity performed and allows for comparisons to be made in the representation as well as revealing

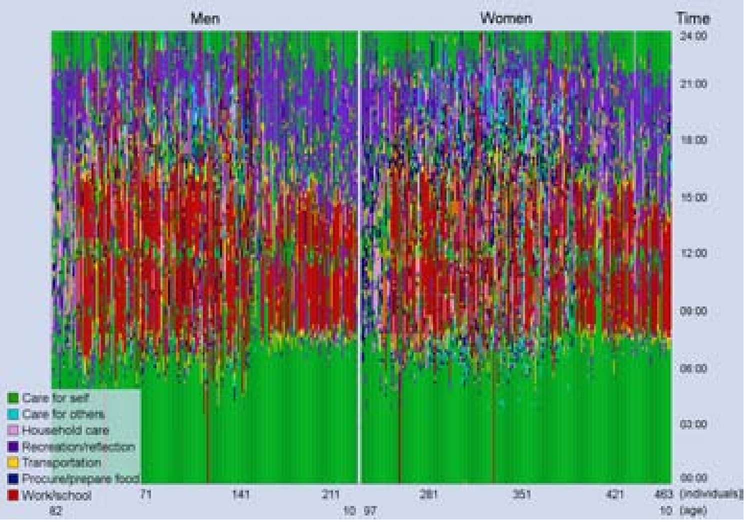


Figure 4.1: Time-geographical representation of VISUAL-TimePAcTS as described in section 3.4.1. A population of individuals aged 10 to 97 is displayed ordered by sex and ascending age from right to left.

initial, general trends at first glance. Sorting the diaries with respect to variables such as sex, age, region and family, makes it possible to further emphasise such trends. In figure 4.1, for example, the diary representation of a population is displayed ordered by sex and ascending age from right to left. Three colour bands representing population activity patterns stand out directly in the display; a sleeping pattern shown in green, a work/school pattern shown in red, and an evening relaxation pattern shown in purple. Even though general population trends become apparent in the representation, specific, more detailed activity patterns also exist within it but these remain very difficult to observe.

Series of single activities or ranges of activities can be highlighted in VISUAL-TimePAcTS and their distribution examined in isolation from surrounding activities but still in the context of the individuals' days. Selecting and highlighting a series of activities reveals the manner and order in which these are combined by individuals during the day, their repetition and their timing. This makes for a first attempt to identify patterns of behaviour in terms of performed sequences. Figure 4.2, for example, displays the same population as figure 4.1 with the activities *'meal'* and *'meal preparation or after-work'* highlighted showing a clear distribution pattern with respect to time of day as well as sex. Looking at when sequences occur in the representation reveals their distribution patterns and allows individuals that perform sequences in a similar way with respect to time of day, duration and number of interactions to be identified, revealing trends in the population's behaviour.

Interaction and filtering techniques have also been implemented in VISUAL-TimePAcTS that allow the analyst to interface with the diary data and retrieve details concerning the

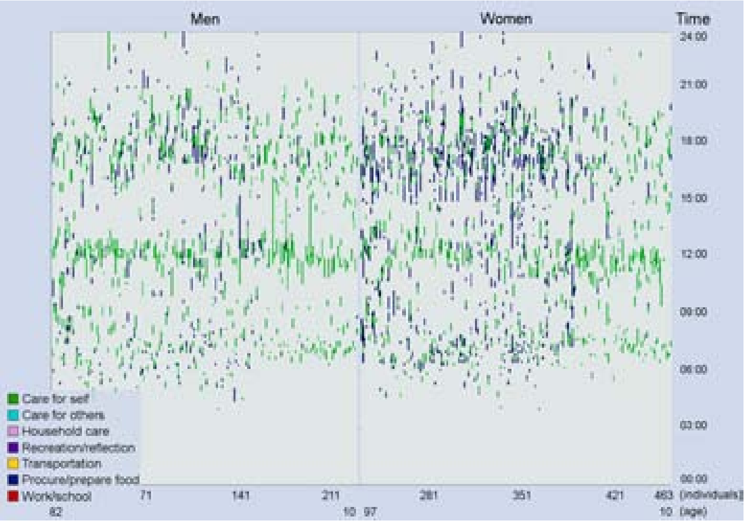


Figure 4.2: Time-geographical representation of VISUAL-TimePAcTS as described in section 3.4.1. A population of individuals aged 10 to 97 is displayed ordered by sex and ascending age from right to left with activities *'meal'* and *'meal preparation or after-work'* highlighted.

activity data records as well as the individuals these correspond to. Finally statistical summaries can be created within the visual analysis tool and exported in formats suitable for traditional time-use analysis. The two analysis methods can, therefore, be combined and a more complete analysis can be achieved.

The visual representation of VISUAL-TimePAcTS, also includes a 3D view of the diaries, showing these as a structure of paths of interchanging activities in an abstract activity space (as described in section 3.4.1). In Paper **II** a controlled user-based evaluation was carried out in order to evaluate this main time-geographical representation implemented in the visual analysis tool, in it's 2D (static 'front' view) and 3D (rotatable view), when exploring the data in search of similarity in individuals' activity paths and, in particular, identifying when individuals perform activities concurrently. The evaluation showed that the 3D view of the application has advantages in performing such tasks for a limited number of individuals since the individuals' activity paths form distinct bundles (groupings) in the activity space, but also showed that when the data size increases cluttering becomes a problem. Furthermore, the interaction mode appropriate for such a representation was discussed in the paper. A constrained rotation around the time axis (y-axis) or shearing along the central z-axis were suggested as more useful interaction schemes for the task in question since they can help avoid disorientation which may occur through a free rotation.

4.1.3 Contributions and conclusions

The main contribution of the work presented this far is the VISUAL-TimePAcTS application itself. In an area such as the social sciences where the use of visualization techniques in general is very limited, the introduction of a flexible and interactive visual analysis tool is a significant contribution with respect to the powerful analysis it facilitates.

VISUAL-TimePAcTS, presented in Paper **I**, and its interactive, three dimensional, time-geographical representation of activity diary data, uncovers the complexity of the data and aids the user in making sense of it. The data is displayed in its original sequential form keeping the focus on an individual level while, at the same time, giving a general overview of the behaviour of a whole group at a population level. The incorporation of appropriate interaction, filtering and highlighting techniques play a significant role in visually enhancing the underlying sequences forming patterns of behaviour as well as their understanding and analysis across the data. Furthermore, VISUAL-TimePAcTS allows users of any experience level to explore and make sense of time-use data of this form.

Paper **II** complements this contribution with a controlled user-based evaluation of the effectiveness and efficiency of 2D and 3D representations in a time-geographic environment for performing a representative task. Also, improvements and dangers within such environments are identified and discussed.

VISUAL-TimePAcTS successfully demonstrates how social science research can be promoted by cross-disciplinary efforts involving computer science and mathematics to create new research and data analysis opportunities. It offers the time-use researcher the opportunity to combine traditional analysis methods with novel interactive ones, get a better understanding of the structure and distribution of the data through visualization techniques and interact with them in a flexible way while getting immediate feedback during the process.

The system provides an effective and intuitive approach to understanding and exploring the data and allows the identification of some of the more apparent sequences through visual observation but, when used with larger data sets or when seeking more complex or more subtle patterns visual inspection, is less effective due to the limitations of the user: their inability to retain large and complex information. This issue has been addressed through the inclusion of algorithmic approaches for the automatic identification of sequences from the diaries and will be presented in the next section.

4.2 Algorithmic sequence mining

Taking advantage of the knowledge gained in the first part of the research, and in order to augment the purely visual approach, an algorithmic sequence mining was introduced in Paper **III** for extracting interesting sequences of activities as patterns from activity diaries. This is an example of how algorithmic refinement can aid in the reduction of data and enable the use of higher level visual representations for communicating the results which are easily accessible to the human vision system. An extended version of the work presented in Paper **III** is available in Paper **IV** which includes a querying approach for filtering the identified resulting patterns as well as a more detailed description of the background and analysis scenarios. To demonstrate the general applicability of the approach in other fields it was, also, successfully applied to the exploration of productivity on a construction site (Paper **V**).

An algorithmic method for the identification of sequences as patterns in activity diary data is important for two main reasons. Such a method (1) extracts and reveals precise sequences as patterns, and (2) it allows for unsupervised exploration of the data. An appropriate visual representation that reveals the sequential nature of the data is a good starting point for an exploration aiming at identifying sequences. Being able to select and highlight specific activities within the representation provides additional visual aid in this identification. Constellations are revealed, this way, but they are not clear since surrounding activities, that may be entirely irrelevant to the sought pattern, are also highlighted and these act as noise in the representation and can actually prevent the detection of interesting relationships (as in figure 4.2). Furthermore, such highlighting assumes that the person performing the data exploration has some initial hypotheses to test, which, in turn, expects them to already be acquainted with the nature of the data and the kind of patterns they may include. Such assumptions, however, limit the potential of an interactive analysis. Employing an algorithmic approach, instead, makes it possible to isolate and extract specific sequences as patterns and display only these as results. Furthermore, since the identification is automatic, it promotes unsupervised (goal-free) exploration which can uncover relationships that would otherwise remain hidden, leading to unexpected discoveries as well as new insights.

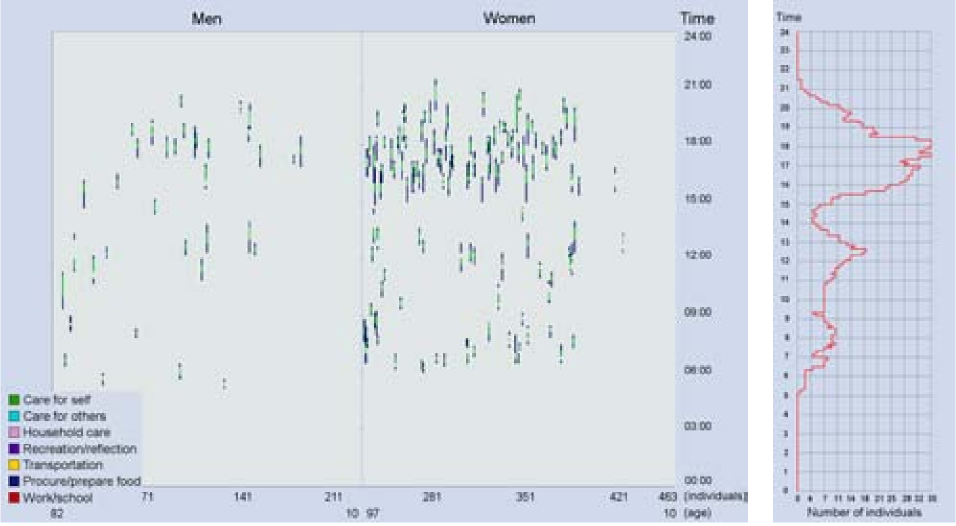
The objective of an algorithmic approach in this context is to automatically identify interesting sequences of activities as patterns in the diary data. The first issue that needs to be addressed in its implementation is concerned with the identification process. Since the goal is to find sequences as patterns, a brute force approach could be used that creates all possible combinations of activities and locates these in the data. This, however, would produce a prohibitively large number of patterns in a prohibitively long time. The problem of constraining such a search, therefore, becomes important.

A second concern in the design of such an algorithm is the definition of what constitutes an interesting pattern. This definition is both domain- and user-dependent and, therefore, more difficult to hard-code in an identification algorithm. Consequently, the implementation should take into consideration which are the characteristics that can make a sequence an interesting pattern, and also allow for user-input in the definition of this interestingness.

4.2.1 Aims

The aim of this second part of the research has been to develop an interactive method that enables automatic identification of interesting activity sequences as patterns. This method should:

- include strategies for reducing the search space, in order to perform the search in interactive times,



(a) The identified sequence seen in the time-geographical representation.

(b) Frequency graph of sequence.

Figure 4.3: Example representation of the results of the sequence mining algorithm. The sequence *meal preparation* → *eat* → *meal afterwork* (wash dishes etc.), identified as a pattern by the algorithm, is being explored. (a) The identified sequence is highlighted in the time-geographical representation showing its distribution in the context of the individuals' days. (b) The frequency graph of the identified sequence is drawn showing its frequency of occurrence during all hours of the day.

- include the possibility to constrain the search so that uninteresting and irrelevant patterns are overlooked,
- allow flexibility with respect to defining interestingness.

User-input should be incorporated in the search that allows the definition of attributes that make a sequence be an interesting pattern. What is interesting, however, is a subjective matter that is best left to the user studying the data to decide. 'Frequent' is not equivalent to 'interesting'. Often the infrequent outliers are what is more important to identify. This is why it is important to allow for user input in the definition of the character of the sequences to be sought.

Finally, after having developed the algorithm in the context of social science activity diaries and after contact with the construction sector, the method was successfully applied to work sampling data in order to study productivity in the building process.

4.2.2 Results

The discussed aims have been addressed through the adaptation and implementation of an Apriori based sequence mining algorithm [3, 4], as described in section 2.2, in VISUAL-TimePAcTS. This work is presented in Papers **III** and **IV** and extended to a second application area in Paper **V**.

Apriori based algorithms have been previously widely used for identifying frequent sequences as patterns in transaction databases. Such algorithms rely on the property that in order for a sequence to be frequent all of it's sub-sequences must also be frequent. Based on this property, patterns are identified iteratively by first constructing candidate sequences as a potential pattern, then locating these candidates in the data and testing the matches against a predefined minimum frequency threshold. The candidate sequences are grown successively, at each iteration, by joining together single activities or previously identified activity sequences. For example, single activities are joined to created candidate sequences of two activities (*2-sequences*), these are then located in the data and their occurrence frequency is tested with respect to the set threshold. The sequences whose frequency count does not satisfy the threshold can be disregarded from any future candidate creation join step, since no super-sequences containing them can meet the threshold. Using the Apriori property in this candidate creation step can lead to a significant decrease in the search space that needs to be considered.

The original frequency criteria of the algorithm were replaced by a number of user-specified limits that can be adjusted in order to define the type of patterns sought and, at the same time, constrain the search. These limits include:

- minimum and maximum sequence frequency,
- minimum and maximum sequence duration,
- minimum and maximum number of data records (diaries) in which the sequence may occur,
- minimum and maximum number of activities allowed to interrupt a matched sequence (size of gap),
- a time window within which a sequence is allowed to occur.

All limit options are incorporated in a graphical user interface within VISUAL-TimePAcTS which enables a user to test different combinations of constraints and assess the results of each test.

The algorithm results in a long list of identified sequences of different lengths. After the identification process has run to completion this list is available to the user who is given the opportunity to click through it and inspect the results. This inspection is performed by displaying the data in the two main representations of VISUAL-TimePAcTS, described in section 3. The identified sequences are highlighted across the time-geographical representation revealing their distribution pattern across the population (figure 4.3(a)), and are displayed in the frequency graph revealing the distribution of their occurrence frequency along the day (figure 4.3(b)). Single or multiple patterns can be represented in the displays. Clicking through the list of results updates the representations which gives the user the opportunity to quickly scan through the list until a distribution of interest appears at which point they can stop and study the pattern more closely by using all the interactive analysis features available in VISUAL-TimePAcTS.

Finally, due to the very large number of identified patterns, a set of filtering commands have been implemented that allow the user to query the list of resulting identified patterns in order to refine the list keeping the interesting ones and filtering out the rest. The implemented commands include logical operators AND and OR as well as a FOLLOWED BY operator which can be applied to single activities, RANGES of activities or a wildcard character (*) denoting that any activity could take that position. All operations can be combined together creating meaningful filtering commands. For example, the pattern list could be queried for patterns revealing how individuals transport themselves to work or school using the command: *'any kind of travel activity' (*RANGE *of activities)* FOLLOWED BY *'work'* OR *'school'*, or for patterns revealing what precedes or succeeds childcare by using the commands *'any childcare activity' (*RANGE *of activities)* FOLLOWED BY *'*'* OR *'*'* FOLLOWED BY *'any childcare activity'*.

4.2.3 Contributions and conclusions

The main contribution of the sequence mining approach presented in Paper **III** and **IV** is the information and relationships that it makes accessible.

All sequences matching some specified constraints are identified. The user is thus given the opportunity to freely explore the distribution patterns of these identified sequences across a population instead of having to query for specific predefined sequences and inspect all occurrences of them disregarding whether these are included in a pattern or not.

Furthermore, this identification and exploration process requires no previous knowledge or initial hypothesis of the user performing the search, apart from optional, general directives. Which, in turn, allows the user to tailor the search depending on personal interest and performed task objective.

Finally, the incorporation of such a data mining algorithm in an interactive visualization environment, such as VISUAL-TimePAcTS, further increases the analysis potential of activity sequences since apart from identifying patterns it also allows a user to study how these appear across a population and make assessments and comparisons with respect to their distribution. The distribution of the pattern *'meal preparation'* followed by *'eat'* followed by *'meal afterwork'* in figure 4.3, for example, reveals a clear concentration amongst women and especially in the evenings. This information would not become apparent without the visual representation linked to the patterns.

The adapted interactive Apriori algorithm implemented in a powerful visualization framework provides a enhanced analysis environment which has been successfully used with two different types of data. It enables a user to extract relationships and convey information that would otherwise remain hidden. Moreover, this combination of techniques allows for insight to be gained concerning the behaviour of a population, for the discovery of unexpected patterns of activity as well as for new hypotheses to arise.

While this semi-automatic algorithmic approach is clearly effective and produces sequences which can be very useful in the exploration of the data, it does have some features which do not fully meet the user's needs. First, the identification is performed in a preprocessing step and the results are then presented for the user to interactively explore. This is a drawback with respect to processing time since candidate sequences have to be created and matched with respect to the specified criteria in the data, which can become a time consuming process. Second, the resulting patterns identified can reach huge numbers that have to be managed, many of these are trivial, obvious or uninteresting, which is a drawback with respect to post-processing time. To address these limitations an interactive sequence identification method was then developed, in the third phase of the research work.

4.3 Interactive sequence identification

An entirely interactive visual data mining system, designed to further enhance the exploration of activity diaries and the identification of sequences as patterns in the data, was developed during the third part of this research. The approach employs a search methodology which takes advantage of the graph structure of the data set and, combined with an intuitive visual interface, enables a user to systematically explore sequences within it. The benefit of this is an effective sharing of the workload between the computer and the user, each performing that task it is better fitted to perform. The work was presented in Paper **VI**.

An interactive visual data mining approach that explores sequences as patterns in real-time presents significant advantages with respect to time and memory load. Most often data mining techniques developed for pattern search perform their identification in an initial pre-processing step. Any visual feedback included in this process usually involves the visual representation and exploration of the algorithm results (similar to the method discussed in section 4.2). This is an effective and useful approach since all the patterns present in the data are identified and made available for the user to asses but, as the data size increases and the constraints relax, the computation time can become prohibitive. Furthermore, any change in the algorithm settings, aiming at refining or altering the direction of the search, requires a full re-computation of the patterns. An interactive approach, instead, can employ a local strategy that systematically explores a single potential pattern at a time. The results are then computed in real-time eliminating the need for an initial global identification step.

Another important advantage of an interactive approach is that it can eliminate the need to store the large numbers of patterns resulting from a traditional pattern mining approach. This need is demanding both with respect to computer memory load, but also with respect to the work load they pose on the user who will scan through them. Considering, moreover, that most of the identified patterns may be obvious or uninteresting to the user, the work load they impose then also becomes unnecessary. An interactive approach, instead, is able to systematically focus the search on the currently explored pattern which, when handling large datasets, provides a more efficient solution. Also, such an approach gives the user the power to decide the direction of their search based on personal interest and task objective.

The development of an interactive visual data mining approach tailored for the identification and exploration of sequences as patterns, poses high demands on the design of the visual interface that will be used to drive it. In order for such a method to work effectively the user should be able to easily understand the structure of the data and the way in which the exploration should proceed. At the same time, the effect and significance of each choice a user makes during this process should be clearly reflected in all representations used for providing visual feedback. If the user is unable to comprehend the chain of events occurring during the process, they will not be able to consciously decide on the direction of the search and the exploration will then become less meaningful. A final challenge lies in the choice of method or algorithm to use for searching and assessing the importance of the explored sequence. Appropriate information should be provided that make it possible to guide the user in interesting directions.

4.3.1 Aims

The aim with this part of the research has been the design and implementation of an interactive visual data mining approach which enables the systematic exploration of activity diaries in search of sequences as patterns. This approach should:

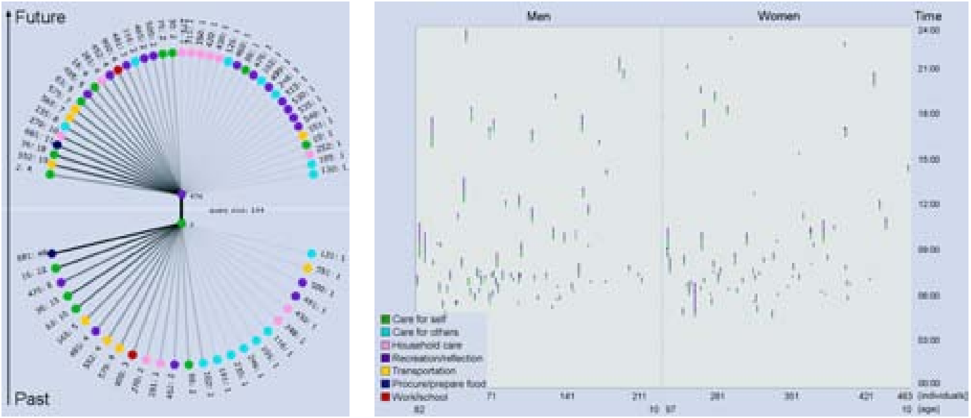
- eliminate the need for pre-processing and post-filtering in the search for patterns,
- take advantage of the inherent structure of the data both in the computation and the representation phase,
- provide immediate feedback concerning the distribution and significance of each explored pattern within the context of the data,
- make use of intuitive and comprehensive visual interfaces throughout the exploration process.

Furthermore, the exploration should be performed in real-time and allow the user to drill-down into the complexity and interconnectivity of the data in a flexible manner.

4.3.2 Results

In order to address the discussed aims the activity diary data have, in Paper **VI**, been treated as a directed graph, similar to a state transition system or the internet, where each activity is considered as a graph node and each transition from one activity to another a directed edge between two nodes. Doing this allows for the use of methods originally created for web search to be used.

The approach that was pursued has its basis in a generalization of the *hubs and authorities* algorithm [40] which was presented in [5] and discussed in section 2.2. This



(a) ActiviTree.

(b) Linked time-geographical representation.

Figure 4.4: The two linked views of the interactive pattern identification environment. (a) 'ActiviTree' visual interface. In the middle the currently explored query-sequence (*'meal'* (code: 2) followed by *'reading'* (code: 476)) is drawn. The activities connecting into the query-sequence are drawn as connecting in-nodes in the bottom of the interface, while the activities connecting out of the query-sequence are drawn as out-nodes in the top of the interface. All nodes are ordered by significance score from left to right. Each node has a label showing the activity code describing it and its frequency of occurrence. Frequency of occurrence is also mapped to the opacity of the edges. (b) Linked time-geographical view showing the distribution of the currently explored query-sequence in the context of the individuals' days.

generalization was adapted in Paper **VI**. In this implementation a single activity or a sequence of activities can be considered as a 'search-key' or query, the *query-sequence*. Given such a query-sequence significance scores are computed for activities connecting into or out of it based on their connectivity information within the activity graph. This query-sequence is created interactively by choosing an initial activity and systematically adding and removing activities to/from it. At each query change the significance scores of all connecting activities are updated facilitating the entirely interactive exploration of activity sequences.

An interactive visual interface was implemented within the VISUAL-TimePAcTS framework for performing this exploration of sequences. A tree-like representation, called 'ActiviTree' (figure 4.4(a)), is used for creating and updating the explored query-sequence and the time-geographical representation of VISUAL-TimePAcTS is used for displaying the distribution of each such explored query-sequence highlighted in the context of the population's diaries (figure 4.4(b)). In the ActiviTree the explored query-sequence is displayed as a sequence of nodes in the middle of the screen, similar to the trunk of a tree, and the activities connecting to the query-sequence are displayed as nodes linking into (*in-nodes*) and out of (*out-nodes*) the query-sequence (as roots and branches to the trunk respectively). The in- and out-nodes are ordered by significance score from left to right in the representation, and their frequency of occurrence as in- or out- nodes is illustrated by the opacity of the connecting edges, completely opaque corresponding to most frequent, and is also printed beside each node. Clicking on a connecting in- ('root') or out-node ('branch') will add this to the query-sequence ('trunk') and update the significance scores as well as the visual representations.

The significance scores of each activity connecting to the query and their frequency of occurrence are cues that reveal information about the connectivity of the data and the next potential exploration step. They aid the user in choosing which activities to add to the query-sequence and, hence, decide on which exploration direction to pursue. The time the algorithm takes to compute the significance scores and draw the ActiviTree at each step is negligible while no activities are overlooked in the process. This allows the user to investigate the distribution of numerous sequences without any restrictions and provides great freedom and flexibility to the exploration.

4.3.3 Contributions and conclusions

Paper **VI** has contributed a visual data mining approach for the systematic interactive identification of significant activity sequences through a flexible exploration interface. The approach gives the user the freedom to decide the direction of the exploration depending on their interest, while at the same time giving information about the significance of the connecting activities at each step of the exploration which can aid them in their choices.

Furthermore, the fact that the sequences are retrieved and explored interactively eliminates the need for pre-processing and post-filtering in the mining process. Due to this the computation time becomes small, the need to store large amounts of results vanishes and the search becomes directed avoiding the retrieval of uninteresting results.

Finally, the approach takes advantage of a user's knowledge, interests and opinions in a tool that balances human interaction and data processing.

Having researched methods concerned with the identification of interesting activity sequences as patterns, a set of interesting supplementary questions arise concerning how these identified patterns actually appear in individuals in a population. Questions such as: What does the way individuals perform combinations of activities say about them? What groupings of performed activity sequences arise? How can similarity with respect to performed activities be measured? Are two individuals that perform similar activities similar in general with respect to their personal characteristics and life status? Such questions brought about the next and final research theme of this thesis, which is concerned with the classification of individuals based on their performed activity sequences.

4.4 Classification based on sequence similarity

The final part of the research implements an interactive visual data mining approach to the clustering of activity diaries based on how these incorporate sequences of activities (activity projects) within them. A set of metrics for assessing similarity between such sequences has been identified and an interactive visualization environment for exploring the clustering results based on this similarity has been proposed. The method has been applied to activity diaries but finds applications in event-based data, in general. The work is presented in Paper **VII**.

Questions arise together with the identification of sequences as patterns in activity diary data as to how these are distributed in a population. Even though interesting sequences can be identified across the diary data, the reason why they appear as they do and the context in which they do, in terms of when, how long, how many times, or with what interruptions, are attributes that are often neglected. Nevertheless, in order to perform a more thorough analysis and better understand the underlying structures that make sequences appear in a certain manner in the diaries, these are attributes that should instead be emphasized. A measurement system that considers these factors is, therefore, important to explore and enables the comparison and determination of similarity between the activity diaries.

Furthermore, since the identification of activity sequences (projects) in individuals' diaries is indicative of their daily behaviour, as discussed previously in section 2.2, a system for measuring similarity based on activity sequences makes it possible to classify a group of individuals' based on this daily behaviour. Moreover, combining this with an interactive visualization system that allows the exploration of the different aspects of the classification results permits even more exhaustive analysis. It facilitates the understanding of why certain groupings occur and opens up the possibility of predicting future behaviour depending on group membership. Within this context it also becomes interesting to examine whether structural similarity between the composition of activity diary records implies an inherent similarity between the owners of the diaries. Whether similar people, with respect to personal attributes such age, sex, education, occupation, income, marital status etc., arrange their days in similar ways. An interface that allows the exploration of meta-information (background information) concerning the clustered diaries makes this possible.

The greatest problem that the implementation of a system for clustering diaries faces is the definition and measurement of similarity between activity sequences, as discussed in section 2.3.1. Since, event-based data, in general, do not have obvious attributes indicating similarity this definition becomes a challenge both with respect to what to measure but also how to decide the level of similarity. Further demands are posed as to how to perform the classification based on this similarity, for example, what algorithm to choose, how to incorporate the similarity measures, how to specify the classification options, and how to include user preference in the process, are all important issues that have to be considered. A final and major challenge concerns the design of an intuitive visualization system that facilitates the fruitful exploration of the clustering results and display of interesting relationships.

4.4.1 Aims

The aim of this part of the research has been to analyse the results of the previously implemented approaches. The goal with the methods presented, so far, has been to identify sequences of activities that suggest interesting patterns of behaviour. Having identified such sequences it becomes interesting to study how they are distributed across a population, how they can be classified into similarly behaving groups and whether these groups display other similar characteristics as well.

Paper **VII** has been concerned with the creation of a visual data mining approach that will cluster individuals based on their similarity with respect to incorporating a given identified sequence into their day. And following that, to interactively explore the resulting clusters along with additional existing information about them in order to further examine their similarity. Such an approach should:

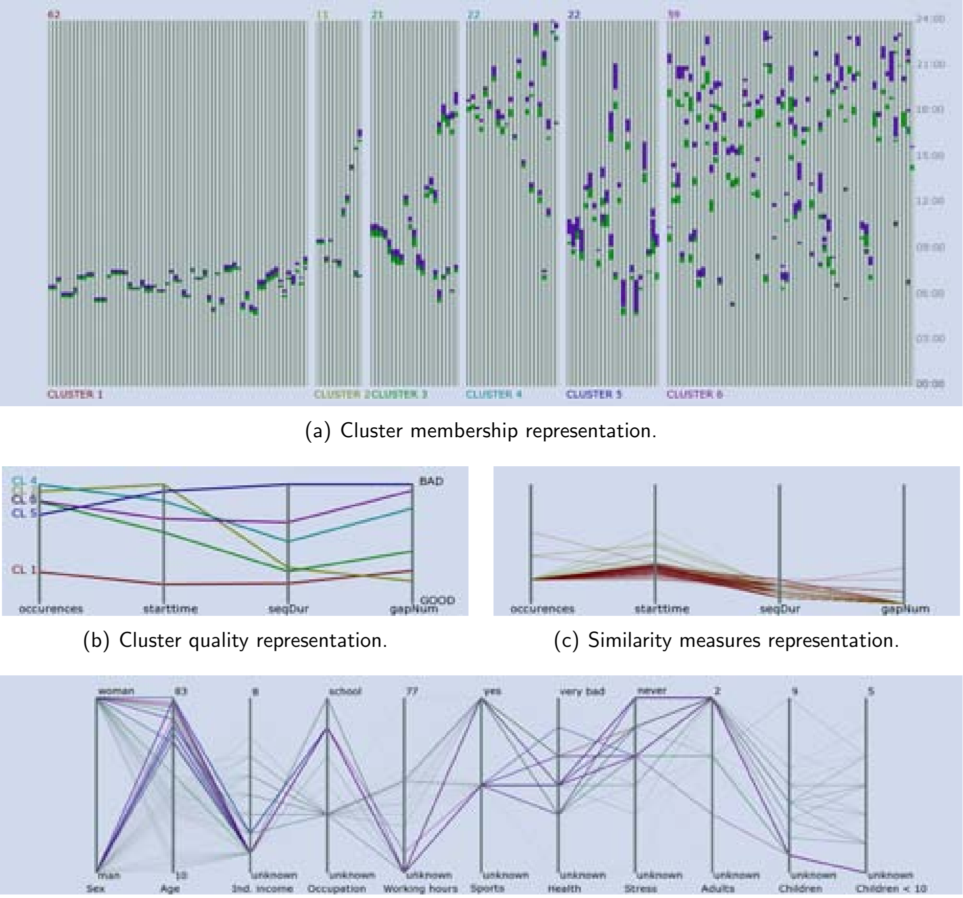
- be based on a set of similarity metrics specific to sequence data,
- allow a user to tailor the definition of this similarity based on the data and task at hand,
- make use of a clustering algorithm that can run at interactive times,
- integrate a set of linked views that allow different aspects of the data and the clustering results as well as relationships between these to be revealed.

4.4.2 Results

An interactive clustering method has been implemented within the VISUAL-TimePAcTS software in order to address the research aims. The method uses previously identified interesting sequences as 'search keys' or 'queries' (*query-sequence*) and locates them in the diaries. These sequences can be identified with any sequence mining approach, including the ones presented in sections 4.2 and 4.3, or can be specified manually by the user.

A set of measures was identified for estimating similarity of how a set of activities appears in the course of the day in individuals' diaries. These measures were extracted based on questions commonly asked in the analysis of event data records and include; When do event-sequences appear? How many times? How long do they last? How many other events appear in between? And how long do these last? Attributes for measuring the general character of an event record were also included such as the amount of fragmentation a record displays, meaning how many events are interchanged between, and the amount of variation present, meaning how many unique events are included in a record.

The identified measures were then used within an interactive clustering algorithm for classifying the identified sequence, the query-sequence, across a population of records. The K-medoids clustering algorithm [31, 82] was applied due to its simplicity, robustness with respect to outliers and empty clusters, and the fact that it considers actual data elements as representative points instead of mean values, this was preferred since the data considered



(a) Cluster membership representation.

(b) Cluster quality representation.

(c) Similarity measures representation.

(d) Background information representation.

Figure 4.5: Linked displays used for representing and interacting with the clustering results. Diaries are clustered *per person* into 6 clusters with respect to the query-sequence *meal* → *reading* and similarity measures: *occurrences*, *start time*, *sequence duration*, *gap size*. (a) Time-geographical inspired cluster representation showing the cluster membership. Each bar corresponds to an individual's diary which are distributed along the x-axis, grouped into clusters with spacing to separate between each cluster. Time is represented on the y-axis going upward. The clusters are ordered by their quality from left to right and the individuals within each cluster are ordered by ascending distance to the medoid from left to right. The query-sequence is highlighted in the context of the individuals' days. (b) Parallel coordinates display showing the *per measure quality* of each cluster, lines represent clusters, each assigned a unique colour, and axes the selected similarity measures. (c) Parallel coordinates display showing the *person-measure matrix* values of each individual in the clusters. (d) Display showing background information concerning the individuals. Each line is an individual and each coordinate axis one of the background attribute values.

in this research are composed of individuals that are optimally left indivisible. The user was given control over all options possible in the clustering process which include the number of clusters sought, the type of metrics to be used in the clustering, the assignment of weighting factors to these metrics in order to prioritize between them, and also whether to cluster the data per data record (individual) or per query-sequence match within a data record. When the second option is chosen the data records are replicated as many times as the given sequence is identified within them.

After the clustering has been performed the results are available for exploration by the user. Four different representations are provided for exploring these results (figure 4.5). A time-geographical representation, similar to the main representation of VISUAL-TimePAcTS, has been implemented to view the clustered diaries' membership. The activity paths of individuals incorporating the query-sequence are displayed along the x-axis with time of day shown on the y-axis, and the query-sequence highlighted within them (figure 4.5(a)) in the context of their days. The individuals in this view are grouped into the extracted clusters ordered by cluster quality from left to right. Within each cluster the individuals are ordered by similarity quality from left to right. Parallel coordinates displays [37] are drawn of the retrieved clusters' qualities (figure 4.5(b)), of the computed similarity measures for each query-sequence match (figure 4.5(c)), and of user selected accompanying background information giving characteristics of individuals' personal and household attributes, as described in section 3.1 (figure 4.5(d)). The user can freely interact with these displays in order to explore the clustering results, for example, by filtering by cluster or by individual and looking at the computed values for each, altering the information displayed in the background information representation and exploring the presence of correlations between personal characteristics and daily behaviour.

The resulting interactive clustering and exploration environment that this approach provides enables the user to better understand the data, retrieve information and relationships that may not be obvious from just examining diaries, and so reach more informed insights about the similarity in the way people act in their daily life.

4.4.3 Contributions and conclusions

The flexible and interactive combination of clustering methods and visualization techniques for analysing event-based data in general and activity diaries in particular, as presented in this section, presents a contribution per se.

The main challenge of this approach is how one quantifies similarity in a valid and representative way. Another main contribution, therefore, is the definition and use of similarity metrics appropriate for clustering event data and in particular activity diary data enabling the identification of similar behavioural patterns.

Furthermore, the analysis of event-based data in terms of similarity is domain- and task-dependent and not always clear, an interactive approach that allows user-input to the definition of this similarity is another significant advantage.

Overall, the combination of tools for steering the algorithm and linked views for displaying and interacting with the results makes for an intuitive exploration interface that enhances the ability of the user to uncover the origin of behavioural similarity.

This final method applied in the presented research has tied together the previously taken steps, taking advantage of methods to identify interesting sequences of activity and further analysing them in order to reveal similarity of behaviour. The definition of similarity, however, may vary depending on the type of data and the objective of the analysis. There is, therefore, more work to be done in refining existing, and adding new similarity metrics for addressing the different questions to be answered.

Furthermore, inherent groupings of the data, such as households, could be considered in the clustering process in order to reveal trends of collective behaviour patterns. These extensions are interesting prospects for future work.

Chapter 5

Conclusions and discussion

The research described in this thesis has been focused on the exploration and analysis of event-based data in general, and the behaviour of sequences within this type of data in particular. Even though the methods implemented and presented are applicable to a wide range of areas the greatest focus and impact has been in the field of social sciences and the analysis of activity diary data. Several contributing approaches have been designed for achieving this and the work has evolved stepwise, each stage building on the previous ones. The knowledge and the identified limitations of one attempt have been the inspiration and motivation for the following one.

This chapter will first summarize the main contributions of the thesis work, presented in chapters 3 and 4, conclusions will then be drawn, and some suggestions for future directions will be discussed.

5.1 Summary of contributions

The main contributions of the presented research work, which have been considered in detail in chapter 4, can be summarized as follows:

- a visual analysis system that facilitates the in-context interactive exploration of activity diaries while preserving their sequential nature in the representation,

- comparative evaluation of the efficiency and effectiveness of a two and three dimensional time-geographical representation in performing the representative task of identifying concurrent events in event-sequence records,

- a method combining algorithmic sequence mining, flexible user input and interactive visualization which allows the identification and exploration of interesting subsequences of events as patterns,

- an interactive environment combining graph similarity notions and visualization allowing the interactive identification and exploration of significant sub-sequences in event-based data,

- the identification of a set of metrics appropriate for measuring similarity between event-sequences based on the distribution characteristics of identified sub-sequences,
- an interactive clustering and exploration environment that enables the classification of event-sequences using these identified similarity metrics.

These contributions are, as mentioned earlier, closely tied together since they represent chapters in a longer learning process. The knowledge gained from each is, therefore, also part of the contributions of this research. Pure visual methods, used initially, are effective in revealing general trends in a population dataset, when aiming at identifying more specific sequences as patterns, however, these are not enough. Automatic methods can be used effectively in order to isolate the interesting sequences from the surrounding activities. The drawback here is the identification time needed as well as the very large number of discovered patterns, many of which are uninteresting. This can be overcome through the implementation of a completely interactive approach to the identification of patterns. Finally, grouping data records based on these identified patterns provides a way to explore whether event-sequences of similar structure also display similarity in other characteristics.

5.2 Conclusions

This thesis has been primarily concerned with visualization and exploration of social science data, in particular the activity diaries used worldwide for performing time-use research. The use of visualization techniques for analysing diary data in this field is limited, however, their analysis commonly involving statistical summaries and graph representations, which conceal the complex character of the collected data. Furthermore, such representations often disregard the individual per se and look at a population through amassed statistics. The impact of the presented work in the field of social sciences and time-use research is significant since it opens a door to the use of sophisticated visualization and interaction techniques. This of course does not dismiss traditional methods, on the contrary the two are complementary and can help the researcher or analyst to go beyond the charts and graphs and into understanding the complexity of the data, making use of more information incorporated in the diaries, exploring the trend and rhythm of daily life and focusing on the role of the individual in it.

The main representation for displaying the activity diaries (event records) consistently present, in all the approaches presented in this thesis, has been a three dimensional one inspired from time-geography, as described in sections 2.1 and 3.4.1. This representation may not be complicated in terms of the computer graphics used to create it, but it is advanced in terms of the information and details it conveys. An evaluation of the effectiveness of this representation, in its two and three dimensional forms, was performed and conclusions could be drawn concerning its performance in carrying out a representative task. The conclusions include that a three dimensional representation can have advantages over a two dimensional one since grouping of data can be identified more easily. Such a representation, however, must be designed very carefully due to potential orientation confusion while interacting with it and can benefit from constraining such interaction.

Visual approaches alone are limited by data size and representation design, as well as how observant a user is. An important conclusion that can be made, therefore, concerns the benefit of automatic identification of patterns in the data. Introducing automation into the identification process can significantly enhance the task. The noise created by surrounding activities is eliminated and the patterns are brought into focus. Most often frequency is specified as the characterizing attribute of an identified pattern. This is, however, not necessarily the desired feature in all cases. Allowing a user to specify the characteristics of what type of patterns to be identified can instead lead to the discovery of patterns that are interesting with respect to other factors apart from frequency. Furthermore, even constraining the identification process by specifying the desired attributes can still lead to a huge number of resulting patterns that match these attributes.

This leads to a subsequent issue which concerns the benefit of introducing interactivity in the search for patterns. Adopting such a method eliminates the need to identify a large number of patterns first and then spend time filtering these patterns in order to retain the interesting ones. Instead this approach allows the user to select and pursue only directions of interest to them during the process.

Interactivity and user input are important features throughout all the research work presented in this thesis. When exploring sequences in event-based data such as activity diaries, the features that are interesting to identify and study depend largely on the task to be performed, meaning the question asked or the hypothesis investigated. Allowing the user to interact with the data and to specify characteristics or constraints to be applied are therefore a necessary flexibility which should not be taken for granted.

Following on from this, what in fact constitutes interesting attributes, in general, is not self-explanatory. The options made available to the user to choose from when exploring a dataset have to be carefully selected and combined taking into account the nature of the data and the type of questions asked during their analysis. Along the same lines, similarity of event records based on how event-sequences are incorporated in them is investigated in paper **VII**. The definition of this similarity is, like the definition of interestingness, not obvious and should also take advantage of the opinion and expertise of the user performing the analysis.

A last conclusion that is important to discuss is the potential use of the methods presented in this thesis. The objective has been the exploration of individuals' daily lives through the study of their day to day activity diaries. The 'big brother' factor of such an exploration is high, since information like this could be used to control, manipulate and invade personal privacy. At the same time, when performed with caution and discretion, the identification of interesting patterns of activity among individuals can have very positive outcomes. It can help understand the way in which individuals structure their days and the reasons for it and can, in effect, help improve their everyday life. To this end, the presented research has contributed a powerful, interactive visual exploration tool for time-use research and analysis of individuals' lives.

5.3 Future work

This research thesis has addressed several aspects concerning the exploration of sequences in event-based data. There are still many interesting future directions that could be pursued.

First, an obvious extension to consider concerning the activity diary data collection is the incorporation of the geographical element in all aspects of the developed methods. Making use of mobile phones and global positioning systems provides accurate location and transportation information. Such information can be included in the representation adding an extra dimension to the visual exploration. Furthermore, geographical position could also be an attribute considered in the filtering, identification and similarity comparison of sequences.

There are many application areas of the presented methods that could be explored and their performance within them evaluated. Incorporating geographical positions expands these possibilities even more. Medical records and the geographical spread of diseases are two important examples. Sequences of medical incidents in individuals could be identified and clustered based on their similarity of occurrence in order to then compare the circumstances of their occurrence and the attributes of the affected individuals. Air traffic control data on flight information is another interesting application area, the identification of patterns of events leading to delays in relation to weather conditions could then be an example task.

The identified metrics for measuring similarity between event-sequences provide a good starting point for comparisons. These need, however, further testing and refinement with respect to the effect of their combination, the appropriate weighting schemes to assign to these combinations, as well as the metrics' applicability and efficiency for various event-based data types. Furthermore, a thorough evaluation of what constitutes similarity in event-based data would be beneficial and could help compose a more complete list of applicable metrics to choose from.

Another area of future interest is that of predicting event-sequence behaviour. Clustering event-records based on sequence similarity permits the probabilistic modelling of event-sequence behaviour within each group. Given then an event-record's group membership the succeeding events can be predicted. Examples of applicability of such an approach are in predicting medical incidents by clustering medical records, in predicting future activity through diary data, and in directed advertising contexts through, for example, internet transaction data.

The exploration of event-based data is an interesting and important topic with a wide application area and enormous potential for future research.

Bibliography

[1] A. Abbott and J. Forrest. Optimal matching methods for historical data. *Journal of Interdisciplinary History*, 16:473–496, 1986.

[2] A. Abbott and A. Tsay. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods Research*, 29(1):3–33, 2000.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. pages 487–499, 1994.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[5] Vincent Blondel, Anahi Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren. A measure of similarity between graph vertices: Applications to synonym extraction and web searching. *SIAM Review*, 46(4):647–666, 2004.

[6] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[7] Lars Eldén. *Matrix Methods in Data Mining and Pattern Recognition.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[8] Kajsa Ellegård. Olikadant. Aspekter på tidsanvändningens mångfald. Occasional Papers 1993:4. Götebors Universitet: Kulturgeografiska Institutionen, 1993.

[9] Kajsa Ellegård. Att fånga det förgängliga. Utveckling av en metod för studier av vardagslivets skeenden. Occasional Papers 1994:1. Götebors Universitet: Kulturgeografiska Institutionen, 1994.

[10] Kajsa Ellegård. A time-geographical approach to the study of everyday life of individuals - a challenge of complexity. *GeoJournal*, 48(3):167–175, July 1999.

[11] Kajsa Ellegård and Matthew Cooper. Complexity in daily life - 3D-visualization showing activity patterns in their contexts. *electronic International Journal of Time Use Research (eIJTUR)*, 1:37–59, August 2004.

[12] Kajsa Ellegård, Torsten Hägerstrand, and Bo Lenntorp. Activity organization and the genration of daily travel: Two future alternatives. *Economic geography*, 53(2):126–152, 1977.

[13] Kajsa Ellegård and Kersti Nordell. *Att byta vanmakt mot egenmakt. Självreflektion och förändringsarbete i rehabiliteringsprocesser.* Johansson&Skyttmo, Stockholm, 1997.

[14] Kajsa Ellegård, Katerina Vrotsou, and Joakim Widén. VISUAL-TimePAcTS/energy use - a software application for visualizing energy use from activities performed. In *Proceedings of the 3rd International Scientific Conference on "Energy Systems with IT"*, Älvsjö, Sweden, 16-17 March 2010.

[15] Kajsa Ellegård and Elin Wihlborg. *Fånga vardagen. Ett tvärvetenskapligt perspektiv.* Studentlitteratur, 2001.

[16] Jerry A. Fails, Amy Karlson, Layla Shahamat, and Ben Shneiderman. A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, October 31 - November 2 2006.

[17] Pip Forer. Timelines environments and issues of risk in health: The practical algebra of (x,y,t,a). In D.J. Briggs, editor, *GIS for Emergency Preparedness and Health Risk Reduction*, pages 35–60. Kluwer Academic Publishers, The Netherlands, 2002.

[18] Tora Friberg. *Kvinnors vardag. Om kvinnors arbete och liv. Anpassningsstrategier i tid och rum.* PhD thesis, University of Lund, 1990.

[19] Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. SPIRIT: Sequential pattern mining with regular expression constraints. In *The VLDB Journal*, pages 223–234, 1999.

[20] Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Mining sequential patterns with regular expression constraints. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):530–552, 2002.

[21] Peter Gatalsky, Natalia Andrienko, and Gennady Andrienko. Interactive analysis of event data using space-time cube. In *Proceedings of the 8th International Conference on Information Visualization (IV'04)*, 2004.

[22] Lise Getoor and Christopher P. Diehl. Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[23] Alex Godwin, Remco Chang, Robert Kosara, and William Ribarsky. Visual data mining of unevenly-spaced event sequences. Interactive Poster in IEEE Symposium on Visual Analytics 2008, 2008.

[24] Torsten Hägerstrand. Statistiska primäruppgifter, flygkartering och "data processing"-maskiner. Ett kombineringsprojekt. *Svensk Geografisk Årsbok*, 33:233–255, 1955.

[25] Torsten Hägerstrand. What about people in regional science? *Papers in Regional Science*, 24(1):7–21, December 1970.

[26] Torsten Hägerstrand. Tidsgeografisk beskrivning - syfte och postulat. *Svensk Geografisk Årsbok*, 50:86–94, 1974.

[27] Torsten Hägerstrand. Survival and arena. On the life-history of individuals in relation to their geographical environment. In Tommy Carlstein, Don Parkes, and Nigel Thrift, editors, *Timing space and spacing time: Human activity and time geography vol. 2*, pages 122–145. Edward Arnold (Publishers) Limited, 1978.

[28] Torsten Hägerstrand. Time-geography. Focus on the corporeality of man, society and environment. *The Science and Praxis of Complexity*, pages 193–216, 1985.

[29] Richard W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160, 1950.

[30] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[31] Jiawei Han and Micheline Kamber. *Data mining. Concepts and techniques.* Morgan Kaufmann Publishers, 2006.

[32] Waqar Haque, Alex Aravind, and Bharath Reddy. Pairwise sequence alignment algorithms: A survey. In *Proceedings of the 2009 conference on Information Science, Technology and Applications*, pages 96–103, Kuwait, Kuwait, March 2009. ACM.

[33] M. A. Hibbs, N. C. Dirksen, K. Li, and O. G. Troyanskaya. Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics*, 6, 2005.

[34] Otto Huisman and Pip Forer. Computational agents and urban life spaces: A preliminary realisation of the time-geography of students lifestyles. In *Proceedings of the third International Conference on GeoComputation*, Bristol,UK, September 1998.

[35] Otto Huisman and Pip Forer. The complexities of everyday life: Balancing practical and realistic approaches to modelling propable presence in space-time. In *Proceedings of the 17th Annual Colloquium of the Spatial Information Research Centre (SIRC)*, pages 155–168, Dunedin, New Zealand, November 2005.

[36] Akihiro Inokuchi, Takashi Washio, and Hiroshi Motoda. An Apriori-based algorithm for mining frequent substructures from graph data. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 13–23. Springer-Verlag, 2000.

[37] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, August 1985.

[38] Chang-Hyeon Joh, Theo Arentze, Frank Hofman, and Harry Timmermans. Activity pattern similarity: A multidimensional sequence alignment method. *Transportation Research Part B*, 36(5):385403, 2002.

[39] Chang-Hyeon Joh, Theo A. Arentze, and Harry J. P. Timmermans. A position-sensitive sequence-alignment method illustrated for space-time activity-diary data. *Environment and Planning A*, 33(2):313–338, 2001.

[40] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:668–677, 1999.

[41] Menno-Jan Kraak. Geovisualization illustrated. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57(1):1–10, 2003.

[42] Menno-Jan Kraak. The space-time cube revisited from a geovisualization perspective. In *Proceedings of the 21st International Cartographic Conference*, pages 1988–1995, Durban, South Africa, August 2003.

[43] Menno-Jan Kraak and Otto Huisman. Beyond exploratory visualization of space time paths. In H.J. Miller and J. Han, editors, *Geographic data mining and knowledge discovery*, pages 431–443. Taylor & Francis, 2009.

[44] Menno-Jan Kraak and Peter Madzudzo. Space time visualization for epidemiological research. In *Proceedings 23rd International Cartographic Conference*, page 302, 2007.

[45] Michihiro Kuramochi and George Karypis. Frequent subgraph discovery. In *Data Mining, IEEE International Conference on*, page 313, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

[46] Michihiro Kuramochi and George Karypis. Finding frequent patterns in a large sparse graph. *Data Mining and Knowledge Discovery*, 11(3):243–271, November 2005.

[47] Mei-Po Kwan. Gender and individual access to urban opportunities: A study using space-time measures. *The Professional Geographer*, 51(2):210–227, 1999.

[48] Mei-Po Kwan. Gender, the home-work link, and space-time patterns of nonemployment activities. *Economic Geography*, 75(4):370–394, October 1999.

[49] Mei-Po Kwan. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: A methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies*, 8:185–203, 2000.

[50] Mei-Po Kwan. GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography*, 86(4):267–280, December 2004.

[51] Mei-Po Kwan and Jiyeong Lee. Geovisualization of human activity patterns using 3D GIS: A time-geographic approach. In Michael F. Goodchild and Donald G. Janelle, editors, *Spatially integrated social science*, chapter 3, pages 48–66. Oxford University Press, 2004.

[52] Heidi Lam, Daniel Russell, Diane Tang, and Tamara Munzner. Session viewer: Visual exploratory analysis of web session logs. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology*, pages 147–154. IEEE Computer Society, 2007.

[53] Bo Lenntorp. *Paths in space-time environments: A time-geographic study of movement possibilities of individuals.* PhD thesis, Royal University of Lund, Dept. of Geography, 1976.

[54] Bo Lenntorp. A time-geographic simulation model of individual activity programmes. In Tommy Carlstein, Don Parkes, and Nigel Thrift, editors, *Timing space and spacing time: Human activity and time geography vol. 2*, pages 162–180. Edward Arnold (Publishers) Limited, 1978.

[55] Bo Lenntorp. The drama of real-life in a time-geographic disguise. In *Sixth Theo Quant Meeting*, Besancon, France, February 2003.

[56] L. Lesnard. Optimal matching and social sciences. CREST (Centre for Research into Elections and Social Trends), January 2006.

[57] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

[58] Heikki Mannila and Pirjo Moen. Similarity between event types in sequences. In *DaWaK '99: Proc. of the First International Conference on Data Warehousing and Knowledge Discovery*, pages 271–280, Florence, Italy, 1999. Springer-Verlag.

[59] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, 1997.

[60] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovering frequent episodes in sequences. In *First International Conference on Knowledge Discovery and Data Mining (KDD'95)*, pages 210–215, Montreal, Canada, 1995. AAAI Press.

[61] Sabrina Mantaci, Antonio Restivo, and Marinella Sciortino. Distance measures for biological sequences: Some recent approaches. *Int. J. Approx. Reasoning*, 47(1):109–124, 2008.

[62] Florent Masseglia, Fabienne Cathala, and Pascal Poncelet. The PSP approach for mining sequential patterns. In *Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 176–184. Springer-Verlag, 1998.

[63] Harvey J. Miller. What about people in geographic information science? *Computers, Environment and Urban Systems*, 27(5):447–453, September 2003.

[64] Kersti Nordell. *Kvinnors hälsa - en fråga om medvetenhet, möjligheter och makt*. PhD thesis, University of Gothenburg, 2002.

[65] Cédric Notredame. Recent progress in multiple sequence alignments. *Pharmacogenomics*, 3(1):131–144, 2002.

[66] Cédric Notredame. Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, 3(8):e123, August 2007.

[67] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Stanford Digital Library Working Papers, Stanford, CA, 1998.

[68] Jian Pei, Jiawei Han, B. Mortazavi-Asl, Jianyong Wang, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. Mining sequential patterns by pattern-growth: The PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440, Nov. 2004.

[69] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-chun Hsu. PrefixSpan mining sequential patterns efficiently by prefix projected pattern growth. In *17th International Conference on Data Engineering*, pages 215–226, 2001.

[70] Donna J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, September 1994.

[71] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. LifeLines: Visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227, New York, NY, USA, 1996.

[72] Catherine Plaisant, Richard Mushlin, Aaron Snyder, Jia Li, Dan Heller, and Ben Shneiderman. LifeLines: Using visualization to enhance navigation and analysis of patient records. In *Proceedings of American Medical Informatics Association Annual Symposium*, page 7680, 1998.

[73] A. Johannes Pretorius and Jarke J. Van Wijk. Visual analysis of multivariate state transition graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):685–692, 2006.

[74] A. Johannes Pretorius and Jarke J. Van Wijk. Visual inspection of multivariate graphs. *Computer Graphics Forum*, 27:967–974, 2008.

[75] Fang Ren and Mei-Po Kwan. Geovisualization of human hybrid activity-travel patterns. *Transactions in GIS*, 11(5):721–744, 2007.

[76] Purvi Saraiya, Chris North, and Karen Duca. An evaluation of microarray visualization tools for biological insight. In *Proc. of the IEEE Symposium on Information Visualization*, pages 1–8. IEEE Computer Society, 2004.

[77] Jinwook Seo and Ben Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35:80–86, 2002.

[78] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[79] James Slack, Kristian Hildebrand, Tamara Munzner, and Katherine St. John. Sequencejuxtaposer: Fluid navigation for large-scale sequence comparison in context. In *Proc. German Conference on Bioinformatics*, pages 37–42, 2004.

[80] Robert Spence. *Information Visualization.* Addison-Wesley Longman Publishing Co., Inc., 2001.

[81] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In Peter M. G. Apers, Mokrane Bouzeghoub, and Georges Gardarin, editors, *Proceedings of the 5th International Conference Extending Database Technology, EDBT*, volume 1057, pages 3–17. Springer-Verlag, 1996.

[82] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[83] Ulanbek D. Turdukulov, Menno-Jan Kraak, and Connie A. Blok. Designing a visual environment for exploration of time series of remote sensing data: In search for convective clouds. *Computers & Graphics*, 31(3):370–379, 2007.

[84] Frank van Ham, Huub van de Wetering, and Jarke J. van Wijk. Interactive visualization of state transition systems. *IEEE Transactions on Visualization and Computer Graphics*, 8(4):319–329, Oct/Dec 2002.

[85] Taowei David Wang, Catherine Plaisant, Alexander J. Quinn, Roman Stanchak, Shawn Murphy, and Ben Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 457–466. ACM, 2008.

[86] Taowei David Wang, Catherine Plaisant, Ben Shneiderman, Neil Spring, David Roseman, Greg Marchand, Vikramjit Mukherjee, and Mark Smith. Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1049–1056, 2009.

[87] Chris Weaver, David Fyfe, Anthony Robinson, Deryck Holdsworth, Donna Peuquet, and Alan M MacEachren. Visual exploration and analysis of historic hotel visits. *Information Visualization*, 6:89–103(15), 2007.

[88] Åsa Westermark. *Informal livelihoods: Women's biographies and reflections about everyday life.* PhD thesis, University of Gothenburg, 2003.

[89] Joakim Widén, Magdalena Lundh, Iana Vassileva, Erik Dahlquist, Kajsa Ellegård, and Ewa Wäckelgård. Constructing load profiles for household electricity and hot water from time-use data - modelling approach and validation. *Energy and Buildings*, 41:753–768, 2009.

[90] Clarke Wilson. Activity pattern analysis by means of sequence-alignment methods. *Environment and Planning*, 30(6):1017–1038., 1998.

[91] Clarke Wilson. Activity patterns in space and time: Calculating representative Hagerstrand trajectories. *Transportation*, 35(4):485–499, 2008.

[92] Krist Wongsuphasawat and Ben Shneiderman. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 27 –34, Atlantic city, NJ, USA, October 2009.

[93] Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. *Data Mining, IEEE International Conference on*, 0:721, 2002.

[94] Amir H. Youssefi, David J. Duke, Mohammed J. Zaki, and Ephraim P. Glinert. Toward visual web mining. In Visual Data Mining Workshop IEEE International Conference on Data Mining, 2003.

[95] Hongbo Yu. Spatio-temporal GIS design for exploring interactions of human activities. *Cartography and Geographic Information Science*, 33(1):3–19, January 2006.

[96] Mohammed J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1/2):31–60, 2001.

[97] Jinfeng Zhao, Pip Forer, and Andrew S. Harvey. Activities, ringmaps and geovisualization of large human movement fields. *Information Visualization*, 7(3-4):198–209, 2008.