MASTER

SleepVis
Visual Analytics for Sleep Diary Analysis

Smets, Anne

Award date:
2021

Department of Mathematics and Computer Science
Algorithms and Visualization

SleepVis: Visual Analytics for Sleep Diary Analysis

Master Thesis

Anne Smets

Supervisors:
prof.dr. Anna Vilanova
MSc. Humberto Garcia Caballero
prof.dr. Sebastiaan Overeem
dr. Tim Leufkens

Final Version

Eindhoven, September 2021

Abstract

A sleep diary is a tool that is used to collect information about the daily sleep-wake pattern of people using a subjective perception of sleep. They are often used in the diagnosis and treatment of sleep disorders. In this project, we focus on data collected using a mobile sleep diary application. Analysis of the data is challenging because we are dealing with heterogeneous event sequence data that contains complex temporal patterns and geographical information. To help sleep experts gain insight into the patterns in the data and formulation of hypotheses about the relation between sleep and external factors, we developed a visual analytics system called SleepVis. It allows the users to explore the sleep event sequences and related derived and self-reported sleep parameters on different time scales. We also propose a way to summarize the event sequence data to provide an overview of the different patterns that are in the data. The utility of the system is demonstrated by a user study and two use cases.

Acknowledgements

First of all, I would like to thank Anna Vilanova for her guidance and valuable time during our weekly meetings. She really kept me motivated to keep improving and bring out the best in this project and me. I would also like to thank Humberto Garcia Caballero for always being available to answer my questions. Without his guidance and tips on the implementation I would not have been able to create such an elaborated system. Furthermore, I would like to thank Sebastiaan Overeem and Tim Leufkens for making the time to provide extensive knowledge on sleep analysis and ideas on the design of the system. I would also like to thank Dirk Fahland for being part of the assessment committee. Moreover, I would like to thank Maarten Zinn for designing and creating the icon of SleepVis. Finally, special thanks my family, friends and boyfriend for their tremendous support during this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Every two years, the Gemeentelijke Gezondheidsdienst of Brabant-Zuidoost studies the experiences of residents in the area of Eindhoven Airport with respect to well-being and possible disturbances. These reports indicate an increasing trend in sleep disruption due to air traffic [3]. Humans have the ability to perceive and react to sounds while asleep. Therefore, aviation noise during nighttime can directly cause sleep disturbances. A study by Kim et al. showed a positive correlation between nocturnal noise exposure and sleep disturbance [17]. Immediate effects of sleep disturbance due to noise include an increasing number of nocturnal awakenings, earlier final awakening or a delayed sleep onset [23]. Sleep disturbance plays a major role in health problems, such as stress, depression and obesity [11].

Air traffic can also indirectly cause sleep problems by causing distress and annoyance due to noise and visible cues, i.e., observing an airplane in the sky during daytime [21]. The desire to gain more insight into the effect of air traffic on sleep causes the demand for a sleep measurement method that is able to assess the sleep of a larger population for a longer period of time.

Over the years, many approaches to measure sleep have been developed. This ranges from EEG measurements to mobile sleep tracking applications. Especially the latter have become highly popular over the last few years. These applications focus on objective sleep measurements using motion detection or heart-rate detection using wearables. However, objective sleep measurements usually do not correlate to the subjective sleep experience. One method that is able to capture the subjective sleep is a sleep diary, which is often used by health professionals in diagnosis and treatment of sleep disorders [4]. This is a tool used to collect information about the daily sleep-wake pattern of people using a subjective perception of sleep. Subjects fill in the sleep diary every morning about the night before and log the times they were asleep. The information about the sleep-wake times, represented by a sequence of sleep events, is analysed through simplification of the information such as calculation of sleep parameters like sleep efficiency, sleep onset and the total sleep time. These derived sleep parameters provide an indication of the quality and quantity of the sleep. However, there might be patterns in the event sequence data that are currently not captured by the derived parameters. Exploration of the data can lead to new insights about potential new relevant parameters and the relation between them. Furthermore, the information that is retrieved from a sleep diary might not only be useful in clinical settings, but also in identifying sleep-wake patterns and sleep quality of a larger group of people where relations are currently unknown. Exploration of the data can potentially lead to new relevant insights and, for example, might reveal trends in sleep behaviours related to geographical location or seasonal changes.

To attain a clearer perception of how air traffic affects the sleep quality around Eindhoven Airport, a study is being performed in which a mobile sleep diary application is used to retrieve information about the sleep of residents in the area of Eindhoven Airport. The first step in understanding the relation between air traffic and sleep is understanding the sleep behaviour of subjects in the data. Analysis of the sleep diary data can help the researchers to better understand the temporal sleep patterns and sleep quality and help them with the formulation of new questions and hypotheses. For example, one could be interested in what the sequence of sleep events looks like when someone reports a low sleep quality, or whether people in a certain area follow a similar sleep pattern. The digital format of the data also allows exploration of sleep parameters on a smaller time scale, such as the sleep efficiency of a subject in a certain hour. This might reveal new potential relevant sleep parameters. Hence, a system is required that can support the users with these sort of tasks.

Currently, most sleep diaries are paper-based and analysed by hand by experts in the diagnosis and treatment of sleep disorders. This paper-based format is time-consuming and obstructs the analysis of bigger populations. An automated system could alleviate these restrictions and reduce the effort to explore growing datasets. There is currently no system that allows the analysis of sleep diary data of multiple subjects on multiple temporal levels in numerous geographical locations. Hence, this thesis will focus on the exploration of the sleep diary data collected through a mobile sleep diary application by means of an interactive visual analytics system.

1.2 Problem definition

The goal of this thesis is to help the researchers in the exploration of the sleep diary data to generate new questions and hypotheses about sleep behaviours in relation to external factors. We need to enable identification of temporal sleep patterns and exploration of derived and self-reported sleep parameters to support the experts. To achieve this goal, we first need to correctly process the sleep diary data and calculate related sleep parameters that are of interest to the user. We have to deal with heterogeneous event sequence data that contains complex temporal patterns and geographical information. This poses a challenge, since there is, as far as we know, no standard visualization technique to explore heterogeneous event data of multiple subjects on numerous time scales and geographical locations. Another challenge is identifying suitable summarization methods for the heterogeneous event data to visualize multiple records.

1.3 Contribution

We present SleepVis, an interactive visual analytics system for sleep diary data analysis to assist sleep experts in the exploration of sleep diary data. The design of the visualizations is described in detail and guided by the *four-levels nested design and validation* model of Munzner [22]. We use different visualization techniques that match the geographical and temporal characteristics of the data. SleepVis supports exploration of the multidimensional space of sleep parameters and sleep event sequence data. Furthermore, we present an aggregation and summarization strategy to analyse the heterogeneous event sequence data. This is based on the Match and Mismatch method, a similarity metric for categorical temporal data proposed by Wongsuphasawat et al. [30]. The usefulness of the system to explore heterogeneous event sequence data with complex temporal patterns and geographical information is demonstrated with multiple use cases and a user study.

1.4 Organization of the report

The remainder of the report is organized in the following way. Chapter 2 presents a detailed domain, data and task analysis. A discussion of relevant literature regarding sleep analysis and visualization and work in other domains with similar data and tasks beyond sleep analysis is presented in Chapter 3. Chapter 4 discusses the process to transform the raw dataset in a suitable format for our goals. Chapter 5 gives an overview of SleepVis. Chapter 6 provides a detailed description of how we use clustering to group similar event sequences. A thorough description of

SleepVis is given in Chapter 7. Chapter 8 presents the evaluation of our solution. We conclude this work in Chapter 9 where we discuss the results, limitations and future work.

Chapter 2

Domain, data & task analysis

This chapter describes the domain analysis which covers the target users and their questions and goals. This is followed by the data abstraction and the task analysis that allows to relate to existing visualization techniques in literature. This approach is based on the *four-levels nested design* and *what-why-how* framework of Munzner [22]. The model proposes a structured way to answer questions such as *why* a task is being performed, *what* data is shown and *how* to identify the most appropriate and effective visual encodings for the problem at hand. Hence, in this chapter we focus on the *why* and *what* aspects.

2.1 Domain analysis

SleepVis will be used by sleep experts and the researchers of a study interested in the relation between sleep and air traffic. Data is collected through a mobile sleep diary application, which is filled in by people living in the area of Eindhoven Airport. The dataset contains information about the sleep-wake times, the geographical location and self-reported sleep parameters. A series of discussions with the sleep experts showed that the aim of the experts is to get more insight into the sleep behaviours in large populations and formulate new questions and hypotheses about these sleep behaviours in relation to external factors.

Currently, most sleep diaries are paper-based and analysed manually by clinicians. They are used to get a general idea about the sleep pattern of a subject. When inspecting the sleep diary of a subject, the experts are interested in the sleep-wake times and numerous sleep parameters. Some of the parameters are reported by the subject and are already included in the data. However, there are also many derived parameters of interest that the experts calculate themselves. Those include, for example, the number of awakenings, sleep efficiency and sleep onset latency. These derived sleep parameters are a simplification of the event sequence data representing the sleep-wake times. Together with the experts, we designed a list of sleep parameters that are often used in sleep analysis. In addition to the standard parameters used in the analysis of the paper-based sleep diaries, we also created a new category of parameters called *fragmentation*. There might be trends in the raw sleep event sequences that are currently not captured by the standard derived sleep parameters. The digital format of the data makes it possible to easily focus on shorter time scales such as one hour and the corresponding parameters like local sleep efficiency. This can potentially reveal new relevant sleep parameters. Table 2.1 presents an overview of the categories and related parameters we have defined.

Parameter	Unit	Definition
Total duration		
Terminal wake time	### minutes	The total amount of minutes between the final awakening and lights on
Total bedtime	### minutes	The total amount of minutes in bed
Total intended wake in bed time	### minutes	The total amount of minutes awake in bed before lights off and after lights on
Total sleep intention time	### minutes	The total amount of minutes from lights off to lights on
Total sleep time	### minutes	The total amount of minutes of sleep
Total time awake in bed	### minutes	The total amount of minutes awake in bed between lights off and lights on
Total time out of bed	### minutes	The total amount of minutes out of bed between lights off and lights on
Wake after sleep onset	### minutes	The total amount of minutes awake between sleep onset and the final awakening
Latency		
Sleep onset latency	### minutes	The total amount of minutes from lights off to first time asleep
Fragmentation		
Local bedtime	### minutes	The total amount of minutes in bed in a certain time period
Local sleep efficiency	##%	(Local sleep time / Local bedtime) ∗100%
Local sleep intention time	### minutes	The total amount of minutes intended to sleep in a certain time period
Local sleep time	### minutes	The total amount of minutes of sleep in a certain time period
Local time awake in bed	### minutes	The total amount of minutes awake in bed in a certain time period
Local time out of bed	### minutes	The total amount of minutes out of bed in a certain time period
Other		
Total awakenings	### awakenings	The total number of awakenings between lights off and lights on
Total out of bed	### out of bed	The total number of times a person was out of bed between lights off and lights on
Total sleep efficiency	##%	(Total sleep time / Total bedtime) ∗100%
Total sleep episodes	### sleep	The total number of sleep episodes

Table 2.1: List of sleep parameters to derive, divided into four categories, i.e., Total duration, Latency, Fragmentation and Other.

Together with the experts, we designed a set of questions about the sleep behaviours in the data in relation to external factors which they would like to answer:

1. What is the difference in reported sleep quality between workdays and weekends, months and seasons?
2. In what area is a low sleep quality reported?
3. Do people in a certain area follow a similar sleep pattern?
4. What sleep patterns indicate the variation between nights on an individual or group level?
5. What is the difference in sleep patterns between workdays and weekends, months and seasons?
6. What sleep patterns are common in nights with a low reported sleep quality?
7. What new parameters can be identified from the sleep diary that describe specific sleep structure of an individual on short (minutes or hours) and long (days or weeks) time scales?
8. To what extent does air traffic directly affect sleep?
9. To what extent does air traffic indirectly affect sleep?
10. How does location influence the relation between air traffic and sleep?

The questions focus on understanding the temporal sleep patterns, derived and self-reported sleep parameters and geographical aspect of the data in relation to external factors such as air traffic. Due to time restrictions, it was mutually agreed with the experts that we only focus on the sleep-related questions (Question 1-7) and disregard the questions about the relation between sleep and air traffic (Question 8-10). Understanding sleep behaviour and the geographical aspect of the data is the first step towards insight into sleep and its relation to external factors such as air traffic.

2.2 Data abstraction

The data we use is collected using a mobile sleep diary application in which subjects keep a daily log containing information about the quantity and quality of their sleep. The resulting dataset consists of two data tables. The first table contains the sleep periods where each row represents one night of a single subject and provides information about the self-reported sleep parameters, location and the timestamps at which the subject turned the lights off and on. A complete overview of the parameters can be found in Table 2.2.

Parameter	**Definition**	**Type**
ID	Unique ID	Ordinal
Subject ID	Unique subject ID	Ordinal
Lights off	Time at which the subject turned off the lights in the evening	Timestamp
Lights on	Time at which the subject turned on the lights in the morning	Timestamp
Sleep quality	Perceived sleep quality	Quantitative
Rested	How well rested after waking up	Quantitative
Latitude	Geographic coordinate	Geographical
Longitude	Geographic coordinate	Geographical

Table 2.2: Overview of the sleep period data.

The second data table contains the sleep-wake times, represented by event sequence data. It contains information about the timestamps at which the subject was either asleep, awake in bed or out of bed. Table 2.3 shows a list of the parameters. Each row in the dataset represents a single sleep events of a subject with corresponding start and end times.

Parameter	**Definition**	**Type**
ID	Unique ID	Ordinal
Subject ID	Unique subject ID	Ordinal
Event	Type of sleep event (sleep/ awake in bed/ out of bed)	Categorical
Start	Start time of the event	Timestamp
End	End time of the event	Timestamp

Table 2.3: Overview of the sleep events data.

The complexity of this dataset is caused by the different parameter types on multiple time scales. The parameters that are part of the event sequences change many times over the night, such as the sleep event of a subject, whereas other parameters, such as sleep quality, only change from night to night. There is not a direct and straightforward way to visualise this data for our purpose.

2.3 Task analysis

In this section, we describe the task analysis to address the questions that we presented in Section 2.1. By transforming the domain specific questions into an abstract form, it becomes possible to find meaningful similarities within other domains and design effective interactions and visual encodings. In the task analysis, we follow the framework and terminology proposed by Munzner [22]. Therefore we describe the domain specific tasks in terms of *actions* and *targets*.

The questions of the experts can be divided into three tasks:

T1 Understand the behaviour of sleep parameters with respect to location and time.
The experts are interested in the behaviour of the derived and self-reported sleep parameters with respect to the distribution over geographic location and time, addressing Question 1 and 2.
In order to understand the behaviour of sleep diary parameters over time and location we first need to *derive* new parameters. We want to *identify* the *distribution* and *locate outliers*. Multiple quantitative parameters will be *compared* to find *correlations*.

T2 Understand the behaviour of the sleep event sequence data with respect to location and time.
External influences on sleep behaviour cannot always be identified by only focusing on the derived and self-reported sleep parameters. Therefore, with this task we focus on analysis and exploration of the sleep event sequences of subjects, also with respect to location and time. This relates to Questions 3, 4 and 5.
The first step in understanding the behaviour of the sleep event sequence data over time and location will be to *derive* new categorical parameters. We will *identify* the *distribution* of categorical parameters and *explore* the event sequences in order to find *outliers* and *trends*.

T3 Understand the relation between sleep parameters and the sleep event sequences.
This task relates to understanding the relation between the self-reported sleep parameters and raw sleep event sequences. It also focuses on potential new relevant sleep parameters to describe trends in the event sequences that are currently not captured by the standard derived sleep parameters. This task addresses Question 6 and 7.
For the last task, we also start by *deriving* new parameters. We want to *locate* sleep parameters that are related to specific patterns in the sleep event sequences. We also focus

on the other way around, to *locate* event sequences that are associated with specific sleep parameters.

Based on the tasks, we define a set of requirements that the design and implementation should satisfy.

R1 SleepVis should display the relation between multiple derived and self-reported sleep parameters, addressing task **T1** and **T3**.

R2 SleepVis should display the behaviour of derived and self-reported sleep parameters on multiple time levels, addressing task **T1** and **T3**.

R3 SleepVis should display the behaviour of derived and self-reported sleep parameters over different locations, addressing task **T1** and **T3**.

R4 SleepVis should display the behaviour of the sleep event sequences, addressing task **T2**.

R5 SleepVis should be scalable to many data elements of multiple subjects, addressing task **T1**, **T2** and **T3**.

R6 SleepVis should be responsive. The system should be able to display the relation between multiple aspects of the data and identify subgroups based on knowledge of the experts, addressing task **T1**, **T2** and **T3**.

In the next chapter, we discuss literature of existing visual encodings and interaction techniques that is related to our work with regards to data and tasks.

Chapter 3

Related work

In the previous chapter, we analysed the *what* and *why* by observing the domain situation and analysing the data and tasks and defined a set of requirements. In this chapter, we discuss relevant work in the field of sleep analysis and visualization that is related to our work. Then, we also focus on work in other domains with similar data and tasks beyond sleep analysis.

3.1 Sleep data visualization

In recent years, many visualization methods have been developed in order to analyse sleep data. When considering sleep data analysis, it is important to make a distinction between the visualization of sleep event sequence data or quantitative sleep parameters. In our case, both types are relevant and we therefore discuss both.

Wallace et al. proposed a radial layout to analyse sleep event sequence data, which emphasizes the cyclic nature of time [29]. The radial heatmap is shown in Figure 3.1. Each day is represented by a circle in which the colors indicate the type of sleep. This approach is useful to detect large patterns in the data. However, it is not suited to explore the data, especially the data in the inner circles, in detail since the circle graph gives more space to the outer circles. Furthermore, a sleep event of x minutes is represented in the inner circle by less pixels than in the outer circle. This makes comparison of exact duration of sleep events between multiple circles difficult.

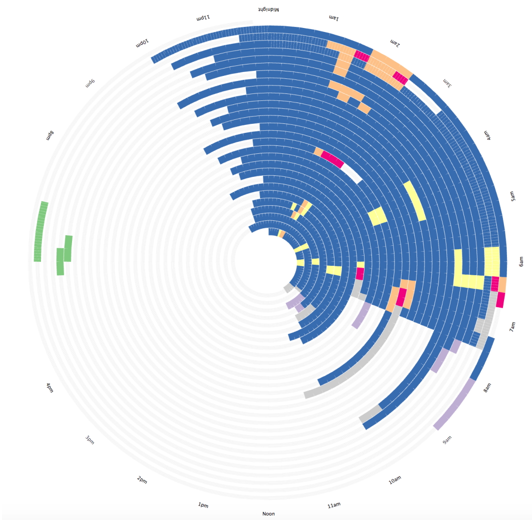

Figure 3.1: A radial representation of temporal sleep event data proposed by Wallace et al. [29].

A more efficient method to visualise the sleep event sequence data is to present the sleep-wake pattern over linear time. A technique that is often used in objective sleep data analysis is the hypnogram, which is used to graphically display the sleep-wake pattern of a subject during the night. A hypnogram can be seen in Figure 3.2. It is often displayed as a line graph displaying three or more categorical sleep stages on the vertical axis over time on the horizontal axis. However, this representation is more suitable for fine time scale, whereas in our case, we are interested in time intervals.

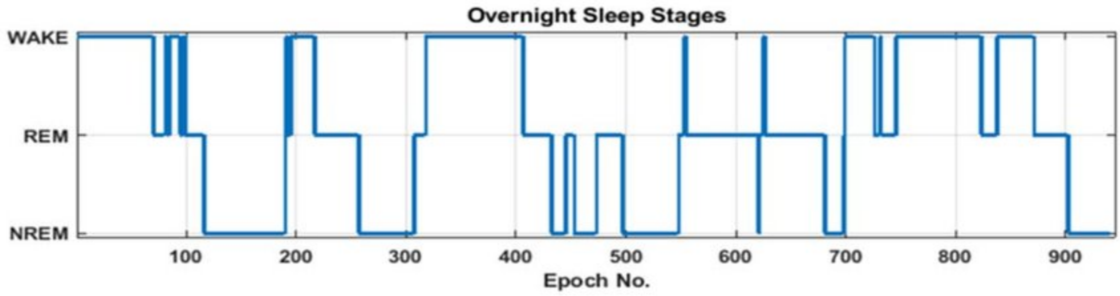

Figure 3.2: A hypnogram to graphically display the sleep-wake pattern [32].

One approach suited for the visualization of time intervals is a timeline. An example where a timeline is used is the system created by Blake & Kerr [4] [5], which is shown in Figure 3.3. It shows a day by day overview of the sleep event sequences of a subject. The colors in the horizontal bars indicate the different events of the subject during the night. This overview is easy to interpret and helps the user to get a clear idea about the sleep-wake times of a subject. Start and end times of sleep events of multiple bars can easily be compared since equal times are mapped to equal positions along the x-axis. Furthermore, sleep events with equal duration are in each bar, as opposed to a radial layout, depicted by an equal number of pixels which allows comparison of duration of sleep events.

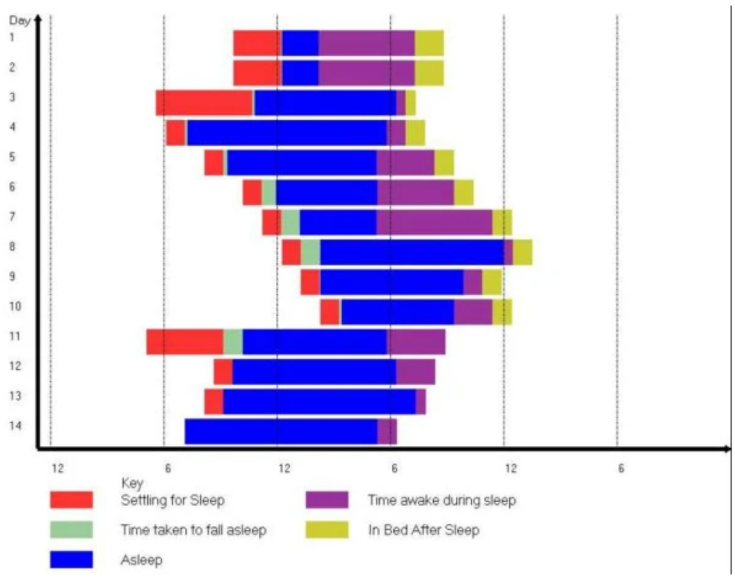

Figure 3.3: A sleep diary graph to provide an overview of the temporal sleep event data [5].

Interaction techniques such as brushing can be used to select a time range of interest. Garcia Caballero et al. [10] demonstrated how brushing can be used on a single timeline to select multiple time ranges, as shown in Figure 3.4. This is useful in case we want to explore the sleep event sequence data and sleep parameters on a small time scale, such as minutes and hours. However, in our case we also need to select time on a higher level such as multiple days or even months. This would require a timeline that ranges over multiple months. Selecting one or multiple days on this timeline will become difficult and would require a high level of precision.

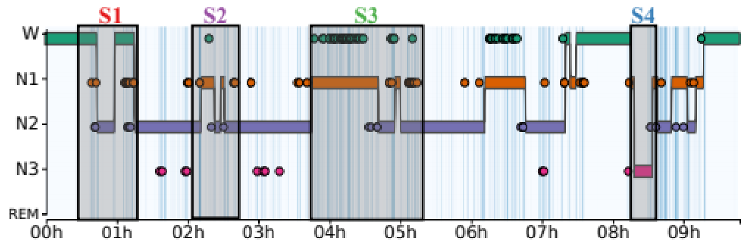

Figure 3.4: A hypnogram where four time ranges are selected [10].

Quantitative sleep parameters are often plotted using standard methods like scatterplots or line graphs. An example of a system that uses these types of visual encodings is the pattern analysis and visualization system (PAViS) developed by Zeng et al. [25]. It aims to monitor sleep patterns of people diagnosed with dementia. This study proposes a visualization system to provide information on the quality, quantity and rhythm of daily, weekly and monthly sleep patterns. They visualise quantitative sleep parameters, such as the total sleeping time, using a simple line chart with time in days on the horizontal axis and sleep quality on the vertical axis. Figure 3.5 shows an example of a monthly overview of the total sleeping time of one subject of one month.

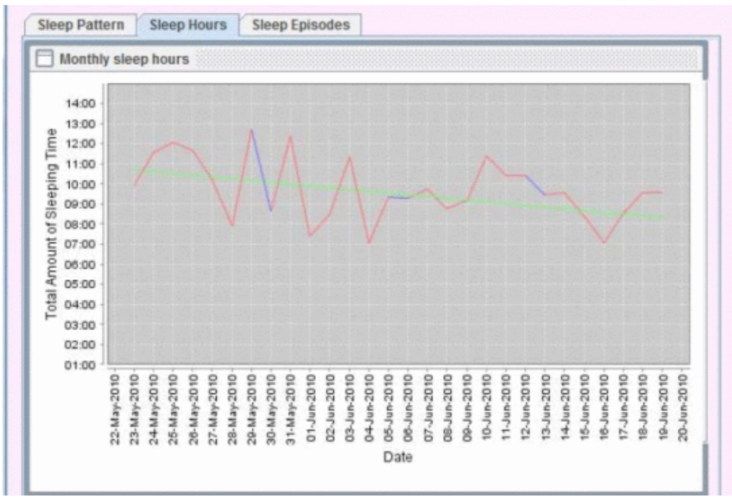

Figure 3.5: A monthly overview of the total sleep hours in PaViS [25].

The timeline and line graph or scatterplot we discussed are well suited to explore the sleep event sequences and quantitative parameters in sleep data of one or multiple nights of a single subject. However, in our case we want to explore data of multiple subjects. We also want to look at multiple sleep parameters at a time to identify correlations to understand the relation between them. Additionally, these techniques do not take other aspects of the data such geographical information into account. Hence, the spatial behaviour of temporal sleep patterns and parameters cannot be analysed. We also discussed how brushing on a timeline can be used to select a time period of interest, especially on a small time scale. This method is less suited to select multiple days or months. In the next section, we look into visualization techniques that are used in other fields beyond sleep analysis for similar multivariate event sequence data.

3.2 Temporal multivariate event sequence data visualization

By abstracting the domain specific questions and data we can find relevant work in other domains that matches our abstracted data and tasks.

A lot of work has been performed in the field of event sequence visualization. Plaisant et al. proposed a visualization system, called LifeLines, to analyse event sequence data from personal and medical records of a single patient [26]. The records are depicted on a graphical time scale. Facets of which the status varies, such as hospitalization or medical problems, are represented by horizontal lines. The color and thickness of the lines indicate the significance of a certain event or relationships between events. Discrete events, such as consultations or medical tests, are illustrated by icons. The drawback of this system is that it is only suited for exploration of data of a single subject and does not support cohort analysis.

Monroe et al. have build on LifeLines by supporting analysis of event sequences of multiple records in a visualization system called EventFlow, shown in Figure 3.6 [20]. Similar to LifeLines, EventFlow represents interval events by lines and discrete events by icons on a horizontal linear time axis. Additionally, EventFlow provides a summary of the records in an aggregated tree-like overview by grouping records with identical event sequences. This tree-based structure branches at the point where event sequences deviate from each other. This summarized representation is based on the icicle plot proposed by Kruskal and Landwher [19]. Events are depicted by color-coded bars, where the color indicates the type of event and the height of each bar represents the proportion of the number of records in that branch with respect to the total number of records. The placement of the bars along the horizontal axis is based on the mean time of the events that belong to the records in that branch. This view is scalable to a large number of event sequences. However, if the number of unique event sequences increases, the view might become cluttered and hard to interpret. Therefore, besides the timeline and aggregated view, filtering and alignment options are offered to enhance the event sequence data analysis. This system only focuses on the analysis of event sequence data and does not provide the ability to explore other data types.

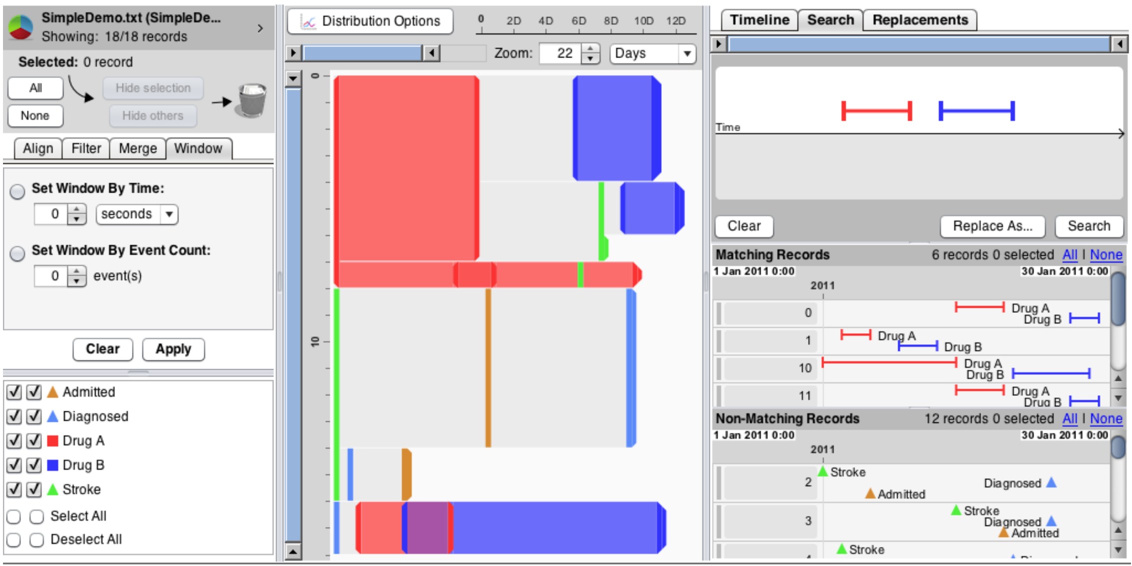

Figure 3.6: An overview of EventFlow, which provides filtering and alignment options, an aggregated view of the event sequence data and a timeline which depicts a detailed view of the event sequence data.

Cappers et al. proposed a system called EventPad to allow simultaneous exploration of event sequences and corresponding parameters of multivariate event sequence data [8]. An overview of EventPad is shown in Figure 3.7. The event sequences are depicted by colored glyphs based on the type of event along the horizontal axis. The distribution of the corresponding parameters of each event in the sequence is displayed by a bar chart in a separate view to explore trends and patterns. The focus of EventPad lies on reduction and analysis of patterns in the event sequences. Hence, the time aspect is omitted in the visualizations. However, in our system, we aim to preserve the time information in the visualizations since time and duration of the events is important in sleep analysis.

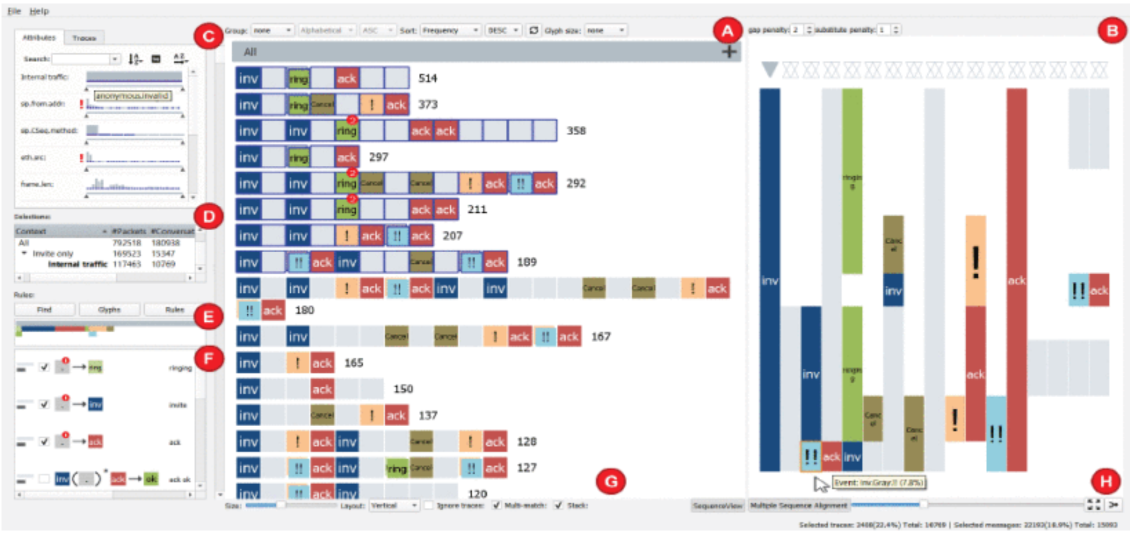

Figure 3.7: An overview of EventPad, a visual analytics system for exploration of multivariate event sequence data.

ST Sequence Miner, proposed by Koseoglu et al., is a visual analytics system to explore event sequence data over time and location [18]. An overview of the system is shown in Figure 3.8. The event sequences are illustrated by colored glyphs along a linear horizontal axis where the color of the glyph indicates the type of event. The spatial aspect of the event sequence data is depicted by means of a cartographic representation. Colored circles on a map show the distribution of events over location. Selections in time, location and event type offer the ability to explore a subset of interest in more detail. The system shows how a combination of multiple views and interaction techniques are used to explore spatio-temporal event sequences. In our case, we are not interested in the spatial aspect of individual events, but of the sequence of events. Furthermore, ST Sequence Miner does not provide the ability to explore the distribution and correlation of parameters belonging to the event sequences.

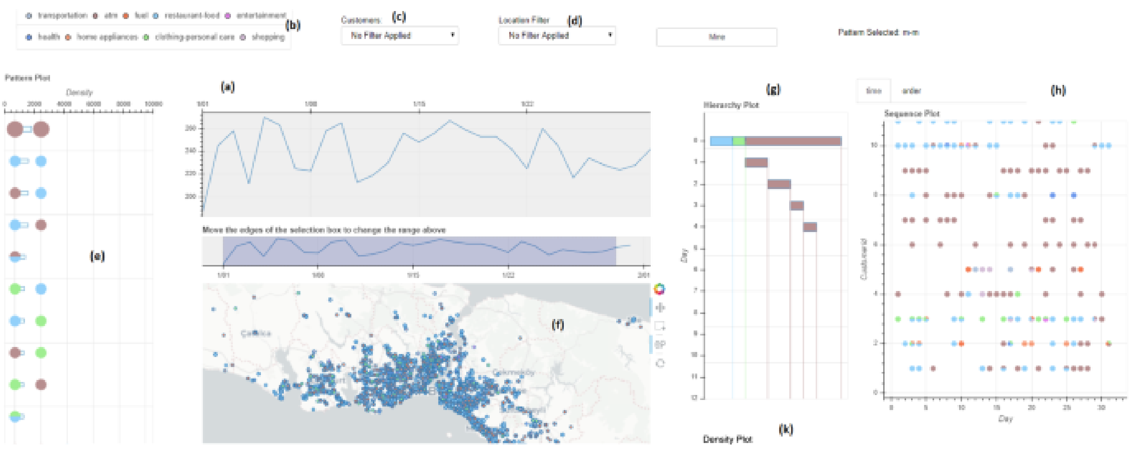

Figure 3.8: An overview of ST Sequence Miner, a system which is used to explore spatio-temporal event sequence data.

Guo et al. presented a visual analytics system to explore multivariate spatio-temporal data called TripVista [12]. They demonstrate how multiple linked views can be used to explore the spatial, temporal and multidimensional aspects of traffic data at a road intersection. Figure 3.9 shows an overview of TripVista. Similar to ST Sequence Miner, TripVista provides the ability to explore the spatial aspect of the data on a map. A ThemeRiver view in combination with glyphs is used to show how the traffic evolves over time. Multiple parameters are depicted by a parallel coordinates plot to explore the multivariate structure of the data. Additionally, interaction techniques, such as brushing in the parallel coordinates plot, enhance exploration of a subset of the data in the other views. In our case, however, we also want to explore the event sequence data in detail, which is not possible with TripVista.

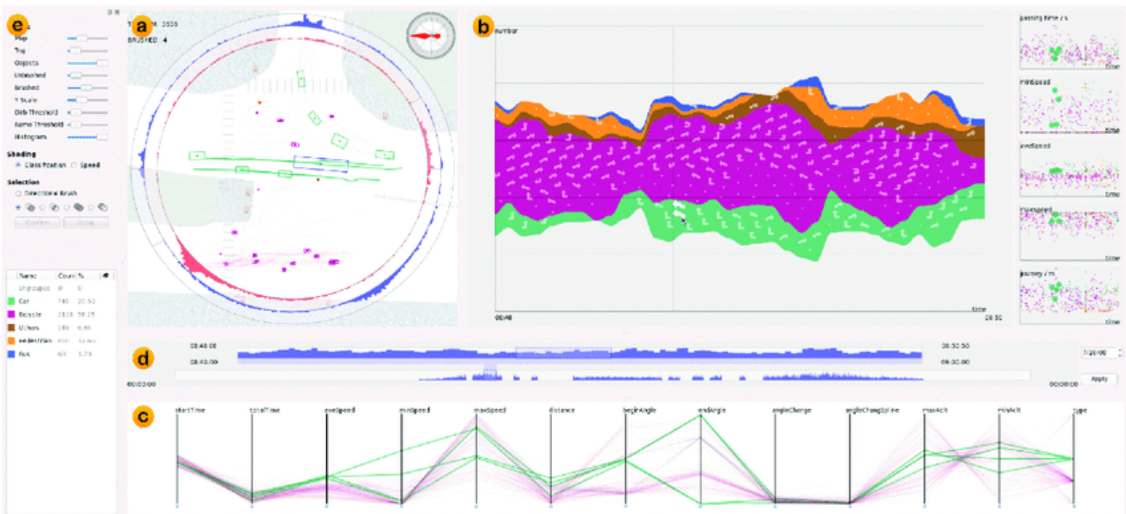

Figure 3.9: An overview of TripVista to explore multivariate spatio-temporal traffic data at a road intersection.

The literature we presented in this chapter shows many visualization techniques suited for visualization of certain parts of our data and tasks. In the domain of sleep visualization we have seen that the focus is mainly on the presentation of data of a single subject. Overall, we can conclude that there is currently no system that enables the option to analyse sleep event sequence data and multiple sleep parameters over time and location of multiple subjects. In the next chapters we will present how we combine multiple visualization techniques to accomplish the tasks and satisfy the requirements we defined in Chapter 2.

Chapter 4

Data processing

In order to answer the sleep-related questions (Question 1-5) and complete the tasks (**T1-T3**) that we defined in Chapter 2, we need to derive the set of sleep parameters that we presented in Table 2.1. These parameters are acknowledged to help experts understand the temporal sleep patterns in the data. In this chapter we discuss the original data and explain how we manipulate the data to calculate these sleep parameters.

4.1 Original data

As discussed in Chapter 2, the original dataset consists of two tables. One table contains the sleep periods, where one row corresponds to one night of one subject. An example of a row is given in Table 4.1.

Parameter	**Value**
ID	5301
Subject ID	4
Lights off	01-06-2021 23:35
Lights on	02-06-2021 08:20
Sleep quality	0.6
Rested	0.7
Latitude	51.336459
Longitude	5.4009012

Table 4.1: An example of a sleep period entry.

The other table consists of the sleep records in which each row contains information about one sleep event. The records that belong to the example row in Table 4.1 are presented in Table 4.2.

Subject ID	**Event**	**Start**	**End**
4	Awake in bed	01-06-2021 23:15	01-06-2021 23:50
4	Sleep	01-06-2021 23:50	02-06-2021 08:00
4	Awake in bed	02-06-2021 08:00	02-06-2021 08:40

Table 4.2: An example of three sleep records entries.

In order to calculate the parameters in Table 2.1, the sleep event records need to be linked to the correct sleep period. In the next section we discuss how we transform the two original data tables into one table.

4.2 Data manipulation

In order to connect the two data tables, we take into account the subject ID, the start and end times of the sleep events and the lights off and lights on markers. The basic idea is that for each row in the periods table we look at the records table to see if there are any events that match the subject ID and where the event falls either within or overlaps with the lights off and or on markers. Algorithm 1 shows how the related sleep events are mapped to the correct sleep period.

Algorithm 1: Map sleep event sequence to correct period.

```
Input: Table of periods P, Table of records R
Output: Table of periods P with added sleep event sequence
for period p in P do
    let A_p be a new array
    for record r in R do
        if p.subject_id = r.subject_id AND
        (r between p.lights_off and p.lights_on OR r overlaps with p.lights_off or p.lights_on)
         then
            append record r to A_p
    Add A_p to period p
return periods P
```

This results in sleep events being mapped to the night of which the event either overlaps with the lights off or lights on markers or falls completely between these markers. Events that do not meet these criteria are not included in the new dataset. Figure 4.1 shows an example of a night where the events that are included are marked with green and the events that are excluded are marked with red. In this example, one *Awake in bed* event and two *Out of bed* events are excluded. However, as the users are interested in sleep in relation to external factors, the focus of the sleep analysis is on the sleep intention times, corresponding to the times between the lights off and lights on markers.

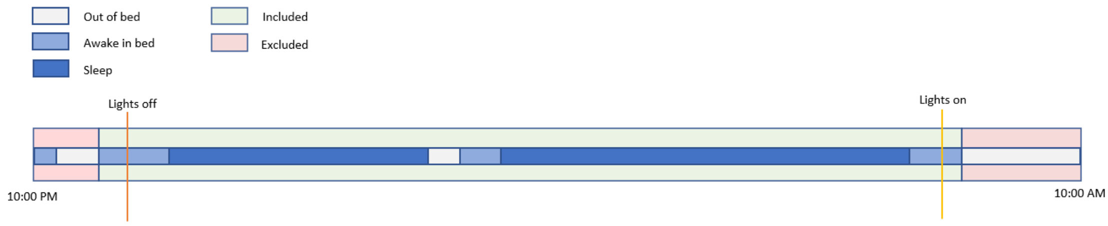

Figure 4.1: Example of a night indicating what sleep events are included (highlighted in green) and excluded (highlighted in red).

We noticed that the way of how the sleep diaries are filled in differs between subjects. Some subjects only fill in events of the night between the lights off and on markers whereas other subjects also include *Out of bed* events during daytime. It can be assumed that people are out of bed when they are not awake in bed or asleep and therefore these *Out of bed* events are redundant. Hence, we decide to discard *Out of bed* events if it is the first or the last event in the sequence of sleep events.

After mapping the sleep events to the correct sleep periods, there are still some periods without any sleep events. This could mean that something is going wrong with transferring the data from the mobile sleep diary application to the data server. Another option is that the subjects did not fill out any sleep events and only reported the sleep quality and how rested they felt after waking up. Since one of the tasks of the user is to understand the behaviour of the temporal sleep patterns, we filter the data such that we only have sleep periods left that contain at least one sleep event.

4.2.1 Sleep events

In order to simplify the calculation of sleep parameters, we decided to aggregate the three sleep events (*Sleep*, *Awake in bed* and *Out of bed*) with the lights on and lights off markers. This resulted in six new sleep event types: *Sleep with lights off, Sleep with lights on, Awake in bed with lights off, Awake in bed with lights on, Out of bed with lights off* and *Out of bed with lights on.* Hence, the sleep events of the example in Table 4.1 and 4.2 are shown in Table 4.3.

Subject ID	**Event**	**Start**	**End**
4	Awake in bed with lights on	01-06-2021 23:15	01-06-2021 23:35
4	Awake in bed with lights off	01-06-2021 23:35	01-06-2021 23:50
4	Sleep with lights off	01-06-2021 23:50	02-06-2021 08:00
4	Awake in bed with lights off	02-06-2021 08:00	02-06-2021 08:20
4	Awake in bed with lights on	02-06-2021 08:20	02-06-2021 08:40

Table 4.3: New sleep events after aggregating the sleep events of Table 4.2 with the lights off marker at 01-06-2021 23:35 and lights on marker at 02-06-2021 08:20.

4.2.2 Sleep parameters

By matching the sleep event sequences to the correct sleep periods and transforming the sleep events, we are able to calculate the corresponding relevant sleep parameters. A complete list of all sleep parameters in which the experts are interested is presented in Table 2.1. These derived sleep parameters are a simplification of the sleep event sequences in the data. They provide an indication of the sleep of a subject. For example, a high value of *sleep onset latency* means that it took long for a subject to fall asleep, which could be an indication of disturbance due to external factors. Furthermore, the sleep parameters that belong to the *fragmentation* category are not part of the standard sleep parameters that are extracted from paper sleep diaries. They provide the ability to describe specific sleep structure on a short time scale such as minutes or hours and relate to external factors. For example, many low *local sleep efficiency* values between 05:00 and 07:00 in the morning might be an indication of sleep disturbance due to external factors. A detailed overview of how the sleep parameters are calculated is presented in Appendix A.

4.2.3 Location

To make sure that the sleep data is not being able to be traced back to an individual subject, the latitude and longitude are converted into the first four digits of the zip code. These four digits are converted back into a latitude and longitude. Hence, we get an approximation of the location of a subject.

However, changing the first four digits of the zip code back to a latitude longitude pair results in many records in the data with the same location. Therefore, we add a random distance to the location within a circle with a radius of 500 meters where the center of the circle is the original location. In order to uniformly, randomly and independently generate the new locations, we first have to convert the radius into degrees. This is done by dividing the radius in meters by 111 111 which equals 1° according to the United States Naval Academy [1].

$$radius_{degrees} = \frac{radius}{111111} \tag{4.1}$$

The next step is to generate two independent uniform random values u and v in the interval [0,1). We use these values to calculate the new latitude and longitude:

$$\text{new latitude} = \text{original latitude} + radius_{degrees} * u^2 * \sin(2\pi * v) \tag{4.2}$$

$$\text{new longitude} = \text{original longitude} + \frac{radius_{degrees} * u^2 * \cos(2\pi * v)}{cos(\text{original latitude})} \tag{4.3}$$

Figure 4.2 shows the new locations, annotated by the blue dots, of four records with the same first four digits of the zip code, indicated by the red dot, after adding a random distance within a radius of 500 meters.

Figure 4.2: Four new locations (blue dots) with the same original location (red dot) after adding a random distance.

Some sleep periods have missing values for the latitude and longitude parameters. These sleep periods belong to subjects who had location services turned off when reporting their sleep in the sleep diary application. Inspection of these subjects showed that most of them also reported nights where the location information was not missing. The missing values could be replaced, for example, by using a technique called mode imputation. This replaces the missing location values by the location where the subject reported most of the other sleep periods. However, analysis of the data reveals that numerous subjects filled in the sleep diary in many different locations. Replacing the missing location values by applying mode imputation could therefore lead to incorrect assumptions about the data. Hence, we have chosen to not replace the missing values, but to leave the values empty.

Chapter 5

SleepVis overview

This chapter presents an overview of SleepVis in which we relate the views to the tasks and requirements we defined in Chapter 2.3.

Figure 5.1: Overview of SleepVis. (1) The correlation exploration component; (2) Subject selection; (3) The map component; (4) The calendar component; (5) Cluster setting; (6) The icicle component; (7) The timeline component.

An overview of SleepVis is shown in Figure 5.1. In order to understand the behaviour of sleep on multiple aspects, we have chosen to place all components on one page. All components are linked to facilitate selection on multiple aspects of the data. The components on the left side of the screen focus on selection of subjects (Figure 5.1.2), the distribution and correlation of derived and self-reported sleep parameters (Figure5.1.1) and their behaviour over location (Figure 5.1.3). Whereas the components on the right side of the screen (Figure 5.1.4/6/7) focus on the sequences of events and the temporal aspect of the data. Below, we introduce the five main views of SleepVis in relation to the tasks and requirements of Chapter 2.3.

- **The correlation exploration component** (Figure 5.1.1)
  This component provides the ability to identify correlations, locate outliers and explore the

distribution of the derived and self-reported sleep parameters, addressing Task **T1** and **T3** and satisfying Requirement **R1**.

- **The map component** (Figure 5.1.3)
  The map component depicts the relation between the location and the derived and self-reported sleep parameters, addressing task **T1** and **T3** and satisfying requirement **R3**.

- **The calendar component** (Figure 5.1.4)
  The calendar component supports the exploration of the behaviour of the derived and self-reported sleep parameters over time, addressing task **T1** and **T3** and satisfying requirement **R2**.

- **The timeline component** (Figure 5.1.7)
  This component provides a detailed view where we can explore the sleep event sequence data of each night in detail to identify temporal trends and outliers, addressing task **T2** and satisfying requirement **R4**. However, this is not scalable to many event sequences. Through selection and filtering in the other components, we can reduce the number of nights.

- **The icicle component** (Figure 5.1.6)
  When many sleep event sequences need to be analysed, the timeline view can become cluttered and overloaded. Therefore, we provide an aggregated view of the sleep event sequence data in the icicle component to identify trends and outliers, addressing task **T2** and requirement **R4** and **R5**.

We want to explore cohorts of data and relate specific sleep patterns to external factors. Therefore, we need an overview of the different patterns in the data. Analysis of the derived and self-reported sleep parameters does not show this relation to external factors. The icicle component provides a way to summarize the event sequence data and detect trends and outliers with respect to the type and order of sleep events in the sequence. However, it does not take the temporal aspect of the data into account. Therefore, we also provide the ability to summarize the event sequence data where we focus on the temporal aspect of the data by grouping similar event sequences by means of clustering. In the next chapter, we go into more detail on clustering the event sequence data.

Chapter 6

Event sequence clustering

Grouping event sequences that are similar to each other allows to summarize the data and get an overview of the different patterns in the data. It provides the ability to explore cohorts of data and look for deviating patterns that show sleep disruption that might be due to external factors. Therefore, we perform a cluster analysis on the sleep event sequences. First, we discuss what clustering methods exist for event sequence data. Then, we describe and justify the clustering method we use to group similar event sequences.

6.1 Clustering event sequence data

Many clustering algorithms exist to group similar data points. However, most of these techniques focus on grouping numerical data points. In our case, we want to guide the users with finding deviating sleep patterns by grouping similar event sequences. Similar event sequences should contain the same type of events and the same events should occur around the same time.

Huang proposed a clustering algorithm to cluster categorical data, called the k-modes algorithm [14]. K-modes is a modification of the often used k-means algorithm [15]. An important aspect is the need of a similarity measure. The k-modes algorithm uses simple matching as a similarity measure. This is defined as the total number of mismatches between two event sequences. Hence, the distance D between two event sequences X and Y is calculated using the following equations.

$$D(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i) \tag{6.1}$$

where

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{if } x_i \neq y_i \end{cases} \tag{6.2}$$

where x_i and y_i are the i-th event of sequence X and Y respectively.

The first step of this algorithm is to randomly select k event sequences and use them as the initial cluster centroids. This is followed by calculating the distance of each event sequence to each of the cluster centroids using the distance measure from equation 6.1 and 6.2. Each event sequence is mapped to the cluster with the minimum distance. The centroid of each cluster is updated by taking the mode of each index i of all sequences that are mapped to the cluster. These steps are repeated until no changes are made in the modes of the clusters.

A disadvantage of this method is that it results in clusters of event sequences with similar types and ordering of events without taking the temporal aspect of the event sequence data into account. In our case, the temporal aspect is important to be able to link specific sleep structures to external factors. Another disadvantage of this clustering method is that the initial centroids, that are chosen randomly, influence the outcome of the algorithm. Furthermore, the number k should be chosen wisely. When selections are made in the data, which leads to less data being passed to the algorithm, k should also be adjusted. Therefore, for each selection in the data, we would need to find the optimal k.

An alternative similarity measure that is suited for calculating the distance between event sequences is called the Match and Mismatch (M&M) measure proposed by Wongsuphasawat et al. [31]. The goal of the measure is to define the similarity between event sequences based on the sequence of events and its temporal aspect. Contrary to the similarity metric of the k-modes algorithm, the M&M method does not fix the mapping of the ith events together. It calculates the similarity between event sequences based on the number of mismatches, represented by the mismatch score, and the time difference between matching events, represented by the match score. Figure 6.1 illustrates two event sequences with the same ordering of events, e.g., (*Sleep with lights off, Awake in bed with lights off, Sleep with lights off*). According to the distance measure of the k-modes algorithm, the distance between the sequences is zero. However, if we want to be able to relate specific sleep disruptions to external factors, we also need to focus on the time aspect. The *Awake in bed with lights off* event in the bottom sequence occurs earlier than in the upper sequence and is therefore, if due to an external factor, probably not caused by the same external factor. The M&M measure returns in this case a distance bigger than zero since it also focuses on the time aspect of the events. Hence, the M&M measure is more suited to identify similar sleep patterns in the event sequence data and relate specific patterns to external factors. In the next section we explain the M&M measure in more detail.

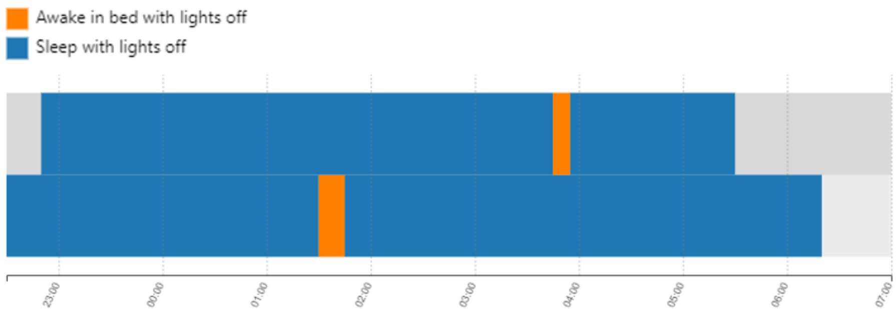

Figure 6.1: Two event sequences that have the same ordering of events.

6.2 Match & Mismatch measure

A sleep event sequence X is a sequence of events (t, c), where t is the start time of an event x with event type c. The index of the event in the sequence is represented by i, such that x_i or (t_i, c_i) is the ith event in the sequence.

$$X = \{(t, c) | t \in Time \text{ and } c \in C\}$$

, where C = {*Sleep with lights off, Sleep with lights on, Awake in bed with lights off, Awake in bed with lights on, Out of bed with lights off, Out of bed with lights on*}.

Suppose we have two event sequences X and Y, illustrated in Figure 6.2, where X is the top sequence and contains the following sequence of sleep events

X = {(23:40, *Awake in bed with lights off*), (23:50, *Sleep with lights off*), (07:25, *Awake in bed with lights on*)}

and Y the bottom sequence where

Y = {(22:50, *Awake in bed with lights on*), (23:05, *Awake in bed with lights off*), (23:10, *Sleep with lights off*), (08:05, *Awake in bed with lights on*)}

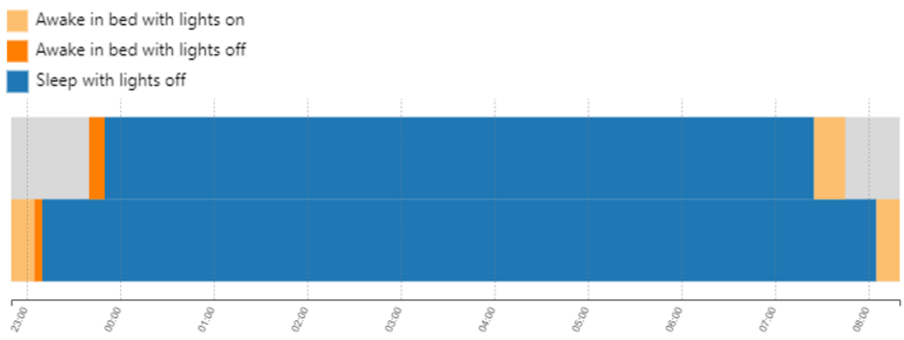

Figure 6.2: Two sleep event sequences X (top sequence) and Y (bottom sequence).

The first step of the M&M measure is to align the event sequences such that they perfectly match with respect to length and frequency of each event category, since this is required by the distance function. This means that the number of events in the sequences should be the same and that there are no pairs that contain different categories of events. Therefore, $(null, null)$ events are added to let the number of events of each event category of the two sequences be equal. Hence, a $(null, null)$ event indicates an extra or missing event.

Sequence X contains three events, whereas sequence Y contains one extra *Awake in bed with lights on* event. Therefore, we add one $(null, null)$ event to sequence X such that sequence X becomes

$X = \{(null, null)$, (23:40, *Awake in bed with lights off*), (23:50, *Sleep with lights off*), (07:25, *Awake in bed with lights on*)$\}$

and sequence Y stays

$Y = \{$(22:50, *Awake in bed with lights on*), (23:05, *Awake in bed with lights off*), (23:10, *Sleep with lights off*), (08:05, *Awake in bed with lights on*)$\}$.

The next step is to calculate the time distance between the events in sequence X and the events in sequence Y using equation 6.3 and 6.4. Since we only want to calculate the difference in time of matching events, we set the distance of non-matching pairs and pairs that consist of two $(null, null)$ events to infinity. The cost for extra or missing events, which are indicated by the $(null, null)$ events, is not included in the time distance function but is managed by the mismatch score which we calculate later.

$$d'((t,c),(u,d)) = \begin{cases} \infty & \text{if } c = null \text{ and } d = null \\ 0 & \text{if } c = null \text{ and } d \neq null \\ 0 & \text{if } c \neq null \text{ and } d = null \\ d((t,c),(u,d)) & \text{if } c \neq null \text{ and } d \neq null \end{cases} \tag{6.3}$$

, where

$$d((t,c),(u,d)) = \begin{cases} |t-u| & \text{if } c = d \\ \infty & \text{if } c \neq d \end{cases} \tag{6.4}$$

For example, the first event of sequence X is a $(null, null)$ event. Therefore, the distance between the first event of sequence X and first event of sequence Y results in zero

$$d((null, null), (22:50, \textit{Awake in bed with lights on})) = 0$$

The second event of sequence X and the first event of sequence Y do not match, which results in the distance being equal to infinity.

$$d((23:40, \textit{Awake in bed with lights off}), (22:50, \textit{Awake in bed with lights on})) = \infty$$

The second events of sequence X and sequence Y match. Therefore, the distance between the events is equal to the difference in time in minutes.

$$d((23:40, \textit{Awake in bed with lights off}), (23:05, \textit{Awake in bed with lights off}) = 35$$

The distance of each pair of events is illustrated in Table 6.1. Where x_1 to x_4 correspond to the events in sequence X and y_1 to y_4 correspond to the events in sequence Y.

	x_1	x_2	x_3	x_4
y_1	0	∞	∞	515
y_2	0	35	∞	∞
y_3	0	∞	40	∞
y_4	0	∞	∞	40

Table 6.1: Table illustrating the distance between each pair of events of event sequences X and Y.

Finally, we calculate the distance between event sequence X and Y using Equation 6.5. We use each value of i and j once and take the sum of event pair distances such that it returns the minimum distance. We use the Hungarian Algorithm to solve this problem.

$$\begin{aligned} D'(X,Y) &= min \sum_{i\in[1,m], j\in[1,n]} d'(x_i, y_i) \\ &= 0 + 35 + 40 + 40 \\ &= 115 \text{ minutes} \end{aligned} \tag{6.5}$$

The final distance represents the difference in time in minutes between two sleep event sequences. We turn this into a normalized score, called the *match score*. If two event sequences are completely equal, meaning the distance is zero, the *match score* will become 1. All event sequences of which the distance is bigger than zero will get a score between 0.01 and 0.99, indicated by Equation 6.6.

$$M(X, Y_j) = \begin{cases} 1 & \text{if } D'(X, Y_j) = 0 \\ \frac{D'_{max} - D'(X,Y_j)}{D'_{max}} * 0.98 + 0.01 & \text{if } D'(X, Y_j) > 0 \end{cases} \tag{6.6}$$

, where D'_{max} is the maximum distance between event sequence X and all other event sequences in the data. Calculated using the following equation.

$$D'_{max} = max_{j\in[1,n]} D'(X, Y_j) \tag{6.7}$$

With the *match score*, we have only focused on the difference in time between event sequences. Hence, we use the *mismatch score* to penalize missing or extra events, indicated by the $(null, null)$ events. To illustrate this, in our example we had to add one $(null, null)$ event to X, which means that Y had one extra event. In order to calculate the *mismatch score*, we calculate the number of mismatches between X and Y, $N(X, Y)$, as the sum of total number of extra and missing events between X and Y. This is also turned into a normalized score, using Equation 6.8 and 6.9, where a *mismatch score* of 1 indicates there are no extra or missing events between two sequences. If the number of mismatches is higher than zero, the score will become a value between 0.01 and 0.99.

$$MM(X, Y_j) = \begin{cases} 1 & \text{if } N(X, Y_j) = 0 \\ \frac{N_{max} - N(X,Y_j)}{N_{max}} * 0.98 + 0.01 & \text{if } N(X, Y_j) > 0 \end{cases} \tag{6.8}$$

, where N_{max} is the maximum number of extra or missing events between sequence X and all other event sequences in the data, indicated by

$$N_{max} = max_{j \in [1,n]} N(X, Y_j) \tag{6.9}$$

Finally, we combine the *match* and *mismatch* scores to calculate the total similarity score $T(X,Y)$ between the sleep patterns with the following equation.

$$T(X, Y_i) = w * M(X, Y_i) + (1 - w) * MM(X, Y_i) \tag{6.10}$$

The total score ranges from 0.01 to 1 where a higher score indicates a higher similarity. The weight w is a value in the interval $[0, 1]$ to give either more importance to the time difference (*match score*) or the number of missing or extra events *(mismatch score)*. In our case, time difference and type of events are both equally important in finding similar sleep patterns. Hence, we use a weight of 0.5.

6.3 Clustering algorithm

The M&M measure returns an asymmetrical similarity matrix of size $n * n$ where n is the number of event sequences in the data. In order to find clusters based on the total similarity scores, we use the Density-based spatial clustering of applications with noise (DBSCAN) algorithm in combination with the M&M measure since this algorithm is able to work with asymmetric distances [9]. Furthermore, as opposed to the k-modes algorithm, it does not require the number of clusters as input.

The main idea of the DBSCAN algorithm is that clusters are areas with a high density that are separated by areas with low density. The algorithm requires two parameters as input: (1) the minimum number of points (*minPts*) in an area to be considered a cluster; (2) epsilon (ϵ) which is the maximum distance between two points to be considered neighbours. DBSCAN groups data points that are within distance ϵ of each other. Data points that do not fall within the distance ϵ of a cluster are labeled as noise. Increasing ϵ will expand the range of a cluster and incorporate more data points into the cluster. If the value becomes too small, clusters can be split into multiple clusters. Since this parameter is dependent on the amount of data, we provide the ability to tweak the value of epsilon value of on demand in the cluster setting component (Figure 5.1.5). For the *minPts* parameter, we use the standard value of 3. If this value becomes too large, the algorithm is not able to detect smaller clusters.

Chapter 7

SleepVis visual design

In this chapter, we present the design of SleepVis. We describe the visual encodings of each individual component in detail and justify our choices. This is followed by a description how the system was implemented.

7.1 Subject selection

The subject selection component (Figure 5.1.2) provides the ability to select one or multiple subjects. Originally, sleep diaries are used in clinical settings where sleep doctors analyse the sleep-wake pattern of an individual subject. This task is still supported by means of this component. However, users can also select multiple subjects to perform a cohort analysis. The component is also useful to exclude subjects from the data that did not fill in the sleep diary correctly such that these deviating sleep behaviours do not influence the analysis.

7.2 The correlation exploration component

The goal of the correlation exploration component (Figure 5.1.1) is to identify the correlation and distribution of the derived and self-reported sleep parameters and locate outliers, which relates to task **T1**.

According to Munzner, a scatterplot is an adequate visual encoding to provide an overview of the distribution, find outliers and also identify the correlation between two quantitative parameters [22]. In order to explore the multivariate data, we need a way to visualise more than two parameters at once. One technique that is often used in the exploration of discrete multivariate data is a scatterplot matrix (SPLOM), where multiple 2D scatterplot charts are organized into one matrix. It is a powerful way to show correlations between multiple parameters, according to Munzner [22]. A drawback of this visual encoding is the scalability and inability to show the multivariate structure of the data. A method that is able to also depict the multivariate structure of the data is a parallel coordinates plot, as shown in TripVista [12]. Multiple axes are placed in parallel next to each other. It is important to note that the the arrangement of the axes is important since this can influence the interpretation of the data. Even though a parallel coordinates plot is more difficult to interpret than a scatterplot matrix, we prefer a parallel coordinates plot to display the multivariate data because it is more scalable than a scatterplot matrix and is able to show the multivariate structure of the data.

Users can select up to five derived or self-reported sleep parameters to analyse together. This maximum number of sleep parameters is needed because more than five sleep parameters would overload the view. Since the list of possible sleep parameters is relatively long, the drop-down provides a search function to help the users with selection of the sleep parameters of interest. Figure 7.1 illustrates how the drop-down provides suggestions based on the input of the user in the menu, to guide the users with selection of a sleep parameter.

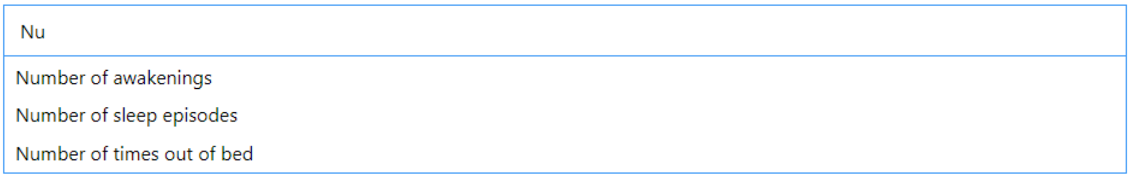

Figure 7.1: The drop-down menu in the correlation exploration component providing three suggestions after the user enters the first two letters of the parameter of interest.

We provide the option to choose between a scatterplot or parallel coordinates plot, depending on the goal of the user. When more than two sleep parameters are selected, the only available option is the parallel coordinates plot.

7.2.1 Scatterplot

When two sleep parameters are selected, the user can select the scatterplot view to identify the relation between the two parameters, supporting requirement **R1**. Each circle in the scatterplot corresponds to one sleep period in the data. More specifically, one night of one subject is represented by one circle. Hence, multiple circles can correspond to one subject. In order to distinguish subjects, the user can hover over a circle, which generates a tooltip showing the subject ID, the cluster ID and the date, as shown in Figure 7.3.

Figure 7.2: Color palette that is used in the correlation exploration component to indicate the cluster.

The color of the circles indicates the cluster ID in order to compare the selected sleep parameters between clusters. The qualitative color palette we use contains 12 distinct colors and is based on ColorBrewer [7], shown in Figure 7.2. According to Harrower et al., increasing the number of colors, such that we get more than 12, will diminish the possibility to clearly distinguish the classes [13]. Since it is possible that we have more than 12 distinct clusters, one color could be used for multiple clusters. However, this happens rarely. To ensure that we can distinguish the clusters when there are more than 12 clusters, the users can hover over a circle to identify the cluster ID.

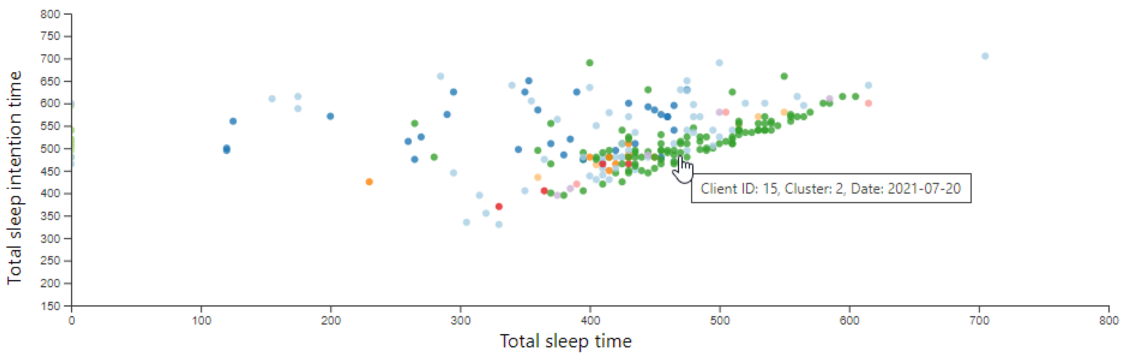

Figure 7.3: Scatterplot showing the relation between *Total sleep intention time* and *Total sleep time*. The tooltip shows the subject ID, cluster ID and date on hover.

7.2.2 Parallel coordinates plot

When more than two sleep parameters are selected, the view automatically switches to the parallel coordinates plot to analyse the relation between five sleep parameters at most, supporting requirement **R1**. Each selected sleep parameter is represented by a vertical axis that ranges from the minimum to the maximum value of the corresponding parameter. Each line between the axes corresponds to one night of one individual. The parallel coordinates plot uses the same color scheme as the scatterplot to color the lines to indicate the cluster.

The arrangement of the axes is important in order to identify the relation between multiple attributes. Hence, the sleep parameters in the drop-down menu can be reordered to change the order of the axes.

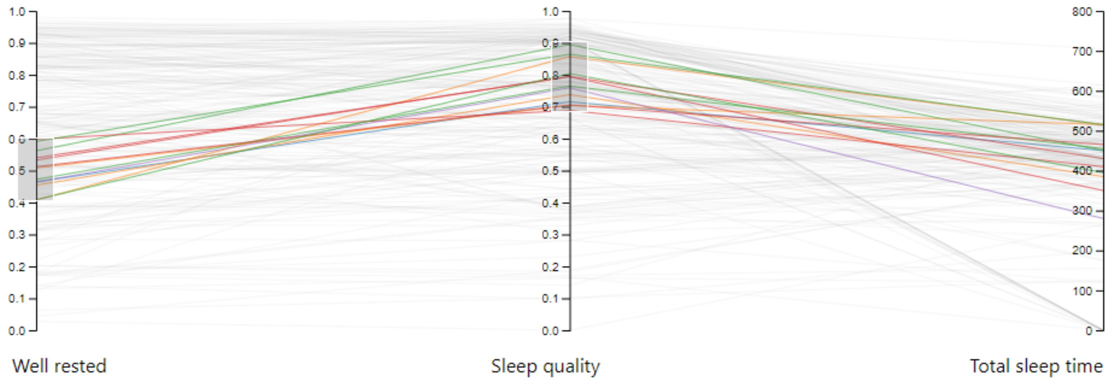

Figure 7.4: The parallel coordinates plot illustrating how the paths that fall outside the brush selection are moved to the background by changing the color, opacity and thickness of the paths.

7.2.3 Average function

In order to easily identify differences in sleep parameters between subjects or clusters, the users can use the average function. This function calculates and plots the average of the selected sleep parameters for each subject or cluster using a single circle or path for each subject or cluster respectively.

7.3 The map component

The main purpose of the map view (Figure 5.1.3) is to depict the relation between the derived and self-reported sleep parameters and the geographical aspect of the data, supporting requirement **R3**. One task is to help the users identify spatial trends and analyse how sleep parameters vary across locations (task **T1**). One technique to depict the distribution of quantitative parameters over location is the combination of a scatterplot and a map, as presented in ST Sequence Miner in Figure 3.8 [18]. We need to take into account that the view can become cluttered if the dataset becomes very large.

Figure 7.5: Multi-hue sequential color scheme that is used to depict the value of sleep parameter on the map.

Each sleep period for which the location information is available is depicted by a circle on the map, as shown in Figure 7.6. The color of the circle corresponds to the value of the derived or self-reported sleep parameter selected in the drop-down menu. Since the sleep parameters are all numerical attributes, we have used a multi-hue sequential color scale from ColorBrewer [7], shown in Figure 7.5. The minimum value of the selected parameter is mapped to the most left color and the maximum value to the most right color of the color scheme.

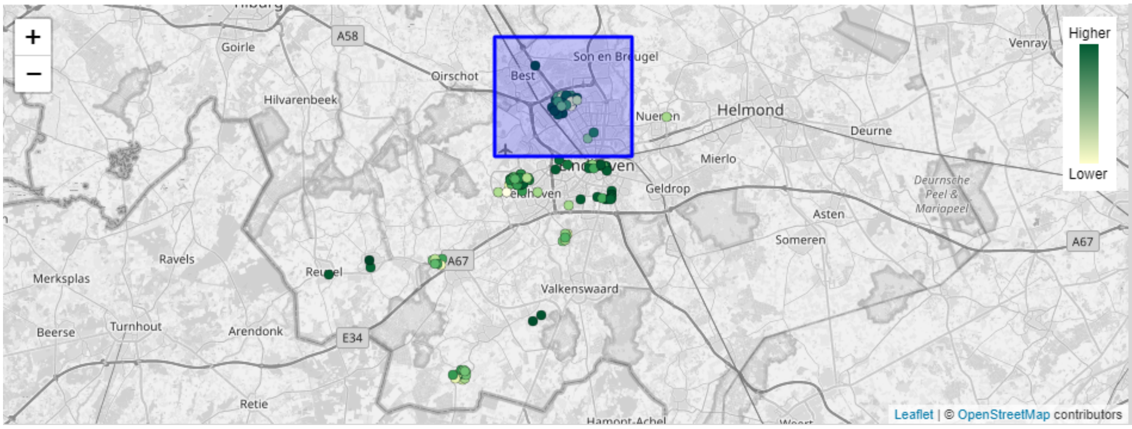

Figure 7.6: The map component in which an area of interest is selected. The selected area is indicated by the blue transparent rectangle.

The map view also supports the function to average the selected sleep parameter per subject or cluster to let the user quickly identify differences between individuals or clusters with respect to location. This is depicted by a circle on the map where the location is the most frequently used location of each subject or cluster. The location is determined by calculating the mode of the zip codes for each individual or cluster and selecting one of the latitude, longitude pairs that correspond to the zip code.

7.4 The calendar component

The goal of the calendar component (Figure 5.1.4) is to provide the ability to explore the derived and self-reported sleep parameters over time, supporting requirement **R2**. A calendar layout is a visual representation that is able to efficiently show temporal patterns. Each day in the calendar can be color coded using an appropriate color scheme to display a quantitative parameter. This approach is used by van Wijk & van Selow to get insight into extensive time series data [28]. This visualization type is a powerful way to express time dependent parameters, such as sleep quality, over time. It supports the identification of daily, weekly or monthly temporal patterns, addressing task **T1**. Furthermore, a calendar layout is useful when the users are interested in selecting time on a higher level. One or multiple days of interest can easily be selected.

The calendar component depicts a calendar heatmap with a sleep parameter of interest. Each day in the calendar is represented by a rectangle. The color of each rectangle corresponds to the average value of the selected derived or self-reported sleep parameter of that day. Higher values are indicated by the darker colors, whereas lower values are indicated by the lighter colors. We use the same multi-hue sequential color scale as in the map component, shown in Figure 7.5. When no data is available, a light grey color is used to increase the attention to the days where data is available. The calendar only shows data of one year because the component would otherwise take up too much space. Especially with the growing dataset. The users can change the year by using the plus and minus button on the left side of the calendar.

Users can select one or multiple days to analyse the sleep patterns and parameters in more detail in the other views to better understand the relation between the sleep patterns and parameters (task **T3**). Selected days are depicted with a red border to clearly indicate the selection without disturbing the exploration and comparison of the sleep parameter of interest, as shown in Figure 7.7. Red is a hue that draws attention and therefore suited to clearly indicate a selection.

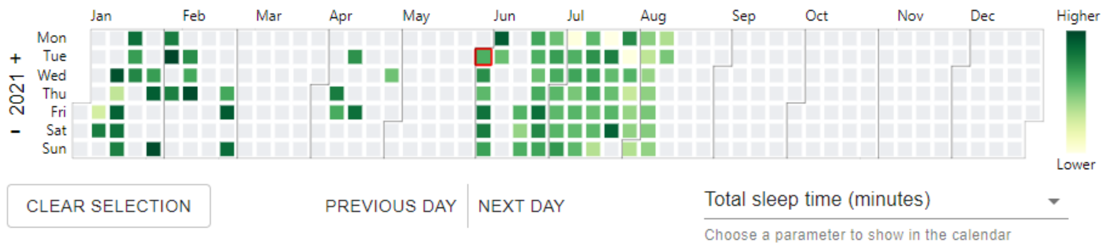

Figure 7.7: The calendar component illustrating the average total sleep time in minutes per day. The 1st of July 2021 is selected, indicated by the red border.

In order to explore the behaviour of sleep parameters and patterns in the other views in more detail per night, users can move to the next or previous day using the next day and previous day buttons respectively. The selection will move to the next or previous day of which data is available, skipping the 'empty' days. This is useful when the users want to explore the sleep behaviour from day to day to find patterns that might indicate disturbance of sleep due to external factors.

7.5 The timeline component

The timeline component (Figure 5.1.7), provides the users a detailed overview of the sleep event sequences in the data, supporting requirement **R4**. A horizontal timeline is used since this method is considered to present an easy to interpret overview of a categorical attribute over time. This also matches the format of the paper-based sleep diaries and will therefore be most intuitive to the users. The focus on the temporal aspect of the event sequence data provides the ability to detect temporal trends and outliers in the sleep event sequences, addressing task **T2**. A drawback of the timeline view is the scalability. Through selections and filtering options in the other components we can reduce the number of event sequences to display.

Colors are used to encode the type of sleep event in the timeline view. The six types of sleep events that we defined in Chapter 4 can be interpreted as three categories where each category has two levels; *lights on* and *lights off*. The sleep events (*out of bed, awake in bed* and *sleep*) are depicted by three different hues. Sleep is associated with a dark blue night sky and therefore encoded with a blue hue. The colors to indicate the other two sleep categories are based on the qualitative color scheme by ColorBrewer shown in Figure 7.2 [7]. The level of a sleep event is depicted by color luminance. We associate lights that are turned off with darkness. Hence, lights off is encoded by a low luminance, whereas lights that are turned on are associated with lightness and therefore encoded by a high luminance. The resulting color scheme is shown in Figure 7.8.

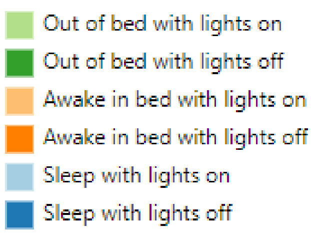

Figure 7.8: Color palette that is used to encode the six sleep event types.

Each event sequence, which consists of one or multiple sleep events, is depicted by a horizontal bar, as shown in Figure 7.9. It is placed along the horizontal axis corresponding to the start and end times of the sleep events and the vertical axis according to the id of the sleep period. The colors of the bar depict the type of sleep event at each point in time.

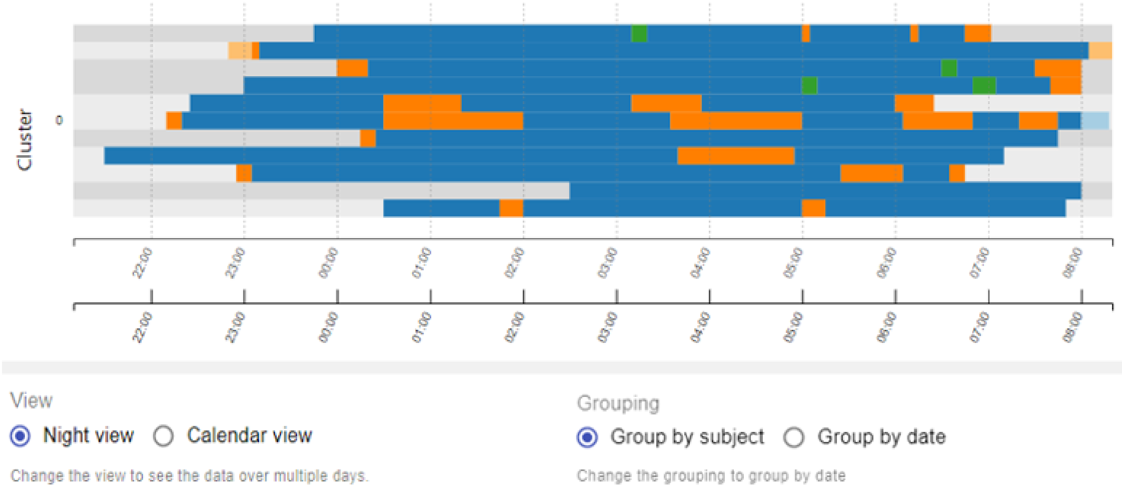

Figure 7.9: The timeline component illustrating the sleep event sequence data when grouping by subject.

To identify the derived and self-reported sleep parameters and location corresponding to an event sequence, users can select an event sequence of interest by clicking on the corresponding horizontal bar. The correlation exploration component and map component highlight the values that correspond to the selection. This is useful in order to understand the relation between the derived sleep parameters and sleep event sequences, addressing task **T3**.

In order to compare the sleep event sequences of multiple subjects, it is important for the user to have a clear indication of what event sequences belong to one subject. Comparison of the sleep behaviours of subjects can provide insight into whether certain sleep structures might be caused by external factors. To identify what event sequences belong to what subjects, we have added the option to group by subject. This option provides grey *swimlanes* in the background of the timeline in order to distinguish the event sequences of different subjects. For each subject, one grey horizontal bar is created. The event sequences are sorted by subject ID, such that all event sequences of each subject are mapped to the correct swimlane. The height of the grey bars corresponds to the number of event sequences of the subjects.

In Figure 7.9, we group by subject and can identify eight swimlanes, which means that the presented data is of eight different subjects. Mousing over a swimlane generates a tooltip that shows the corresponding subject ID.

For the purpose of getting more insight in the patterns and trends in the sleep event sequences and possible external influences on sleep (task **T2**), it is useful for the users to compare the event sequences of multiple days. Hence, users can choose the option to group by date. This sorts the event sequences by date and displays one swimlane for each day. This allows the users to distinguish the event sequences that belong to different days and make comparisons. Mousing over a swimlane produces a tooltip that shows the corresponding date. The same selection of event sequences of Figure 7.9 is shown in Figure 7.10 when grouped by date. The Figure shows two swimlanes, which indicates that the event sequences belong to two different days.

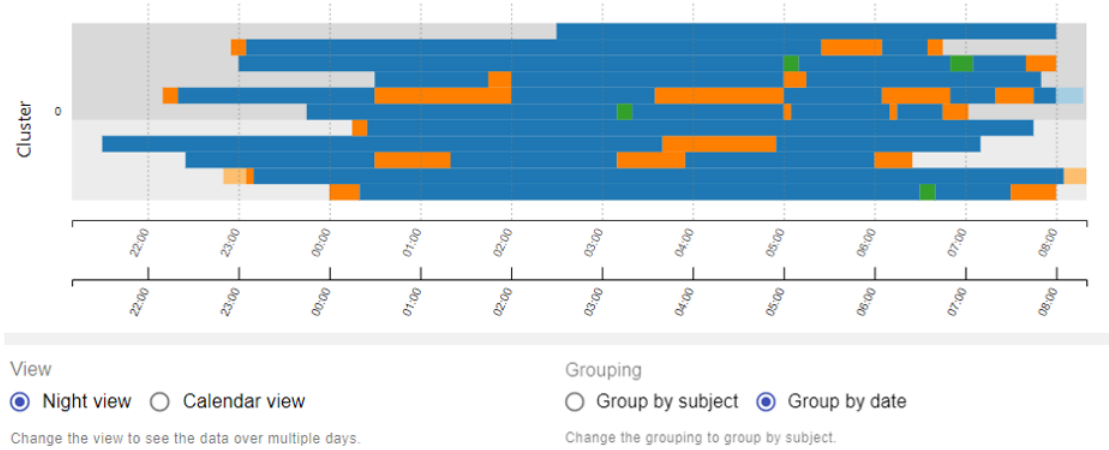

Figure 7.10: The timeline component illustrating the sleep event sequence data when grouping by date.

Additionally, users can change the view option to *calendar view* to compare sleep event sequences of two consecutive nights, as shown in Figure 7.11. The *night view* is selected by default and stacks all event sequences on top of each other, only taking into account time in hours and minutes while ignoring the date. Hence, this allows the users to focus on the differences in time and duration of trends in the event sequences. However, this approach is not scalable to many event sequences since the height of the bars becomes very small, making comparison hard. Therefore, the *calendar view* considers the date aspect of the start and end time of the sleep events and places the sleep periods along the horizontal axis correspondingly. The vertical position is determined by the subject ID. This view is less effective to directly compare time and duration of the sleep periods and more useful if we want to compare the type and number of sleep events in a large number of event sequences of two consecutive days.

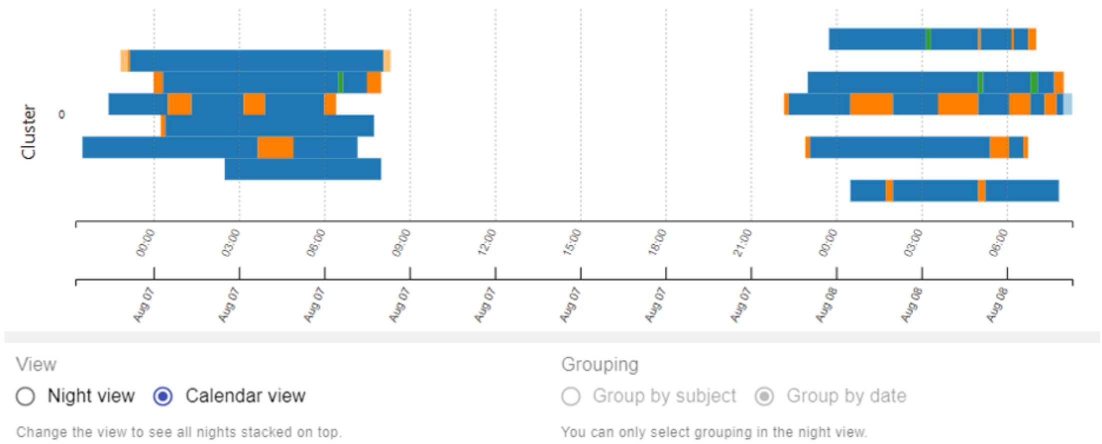

Figure 7.11: The timeline component depicting the sleep event sequences of two consecutive days using the calendar view option.

With the aim to explore the event sequence data to identify potential new sleep parameters that might be relevant to describe patterns in the event sequence data (task **T3**), we provide the option to select a time range of interest by brushing over the horizontal time axis below the timeline view. This is illustrated in Figure 7.12. The selection is linked to the correlation exploration component, which allows the users to explore the local derived sleep parameters that belong to the *fragmentation* category in Table 2.1.

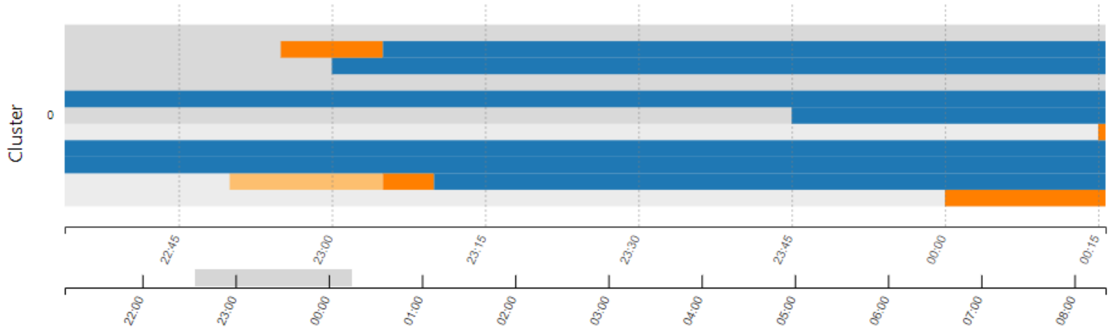

Figure 7.12: The timeline component illustrating how brushing can be used to select a time range of interest.

7.6 The icicle component

The icicle component (Figure 5.1.6) provides a way to summarize the sleep event sequences in the data, supporting requirement **R5**. This representation is inspired by EventFlow, shown in Figure 3.6, in which a summary of event sequences is presented as a tree-like overview [20]. The icicle view gives the users insights into the behaviour of sleep event sequences and allows identifying patterns, addressing task **T2**. The component focuses on the order of the sleep events in a sequence and number of occurrences of a sleep event without taking the time aspect into account. The aggregation of the sleep event sequence data might show deviating patterns that are of interest, such as event sequences where subjects have been awake in bed the whole night.

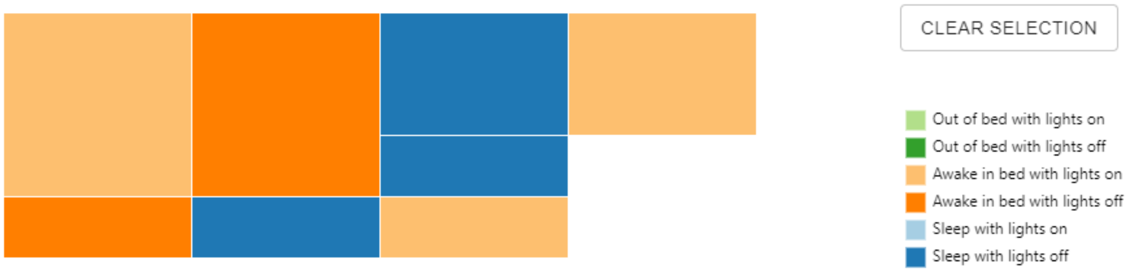

Figure 7.13: Icicle plot showing an aggregation of four sleep event sequences.

The sleep events are depicted by rectangles with equal width. The height of a rectangle indicates the proportion of occurrences of a certain sleep event at index i in all event sequences. For example, Figure 7.13, shows how the icicle plot depicts the the aggregation of four event sequences. Three out of four event sequences start with the event *Awake in bed with the lights on*, indicated by the light orange rectangle on the left side. This event in these sequences is followed by *Awake in bed with lights off* events. One of these three event sequences has *Sleep with lights* as the last event, whereas the other two event sequences also include an additional *Awake in bed with lights on* event. Only one of the sequences starts with the *Awake in bed with lights off* event.

To get a better understanding of the sleep event sequences and the relation to the derived and self-reported sleep parameters (task **T2** and **T3**), users can select an event sequence of interest. The selected sequence will be highlighted, as shown in Figure 7.14 in which the following event sequence is selected: *Awake in bed with lights on, Awake in bed with lights off, Sleep with lights off, Awake in bed with lights on.* The other views are updated accordingly to allow the exploration of the associated details and derived and self-reported sleep parameters of the selected event sequence.

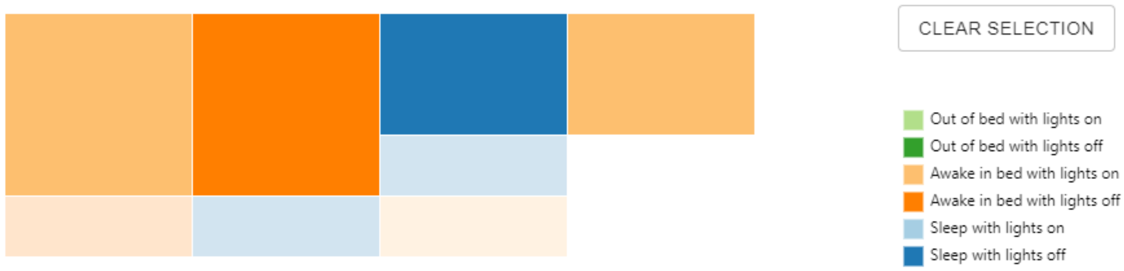

Figure 7.14: Icicle component with selected sequence of events: *Awake in bed with lights on, Awake in bed with lights off, Sleep with lights off, Awake in bed with lights on.*

7.7 Cluster setting

The goal of the cluster setting component (Figure 5.1.5) is to help the users to identify trends and outliers in the sleep event sequences through aggregation, addressing task **T2**. When many event sequences need to be analysed, the Icicle and Timeline component can become cluttered and overloaded. This obstructs the identification of patterns and outliers in the event sequence data. Therefore, the cluster setting component provides the ability to group similar event sequences, supporting requirement **R5**. As discussed in Chapter 6, the clustering algorithm requires epsilon as an input parameter. It is defined as the maximum distance between two points to be considered neighbours. The value of epsilon depends on the level of data and can be changed using the slider that is shown in Figure 7.15.

Figure 7.15: A slider to change the value of epsilon in the cluster setting component.

The results of the clustering algorithm can be interpreted in the correlation exploration, icicle and timeline components. For example, the scatterplot in Figure 7.3 provides insight in the distribution of the derived sleep parameters of each cluster by color coding the circles based on the cluster ID.

The icicle component presents a summarized overview of the event sequences that belong to a cluster by illustrating one icicle plot for each cluster. This allows the users to compare the event sequences of the clusters, focusing on the type, ordering and number of occurrence of events, without taking time into account. Figure 7.16 shows the icicle component when the cluster algorithm returns two clusters. Each icicle plot is placed along the vertical axis according to the cluster. We clearly see that almost all event sequences that belong to cluster 1 start with *Awake in bed with the lights off* whereas for cluster 0, this holds for approximately half of the event sequences.

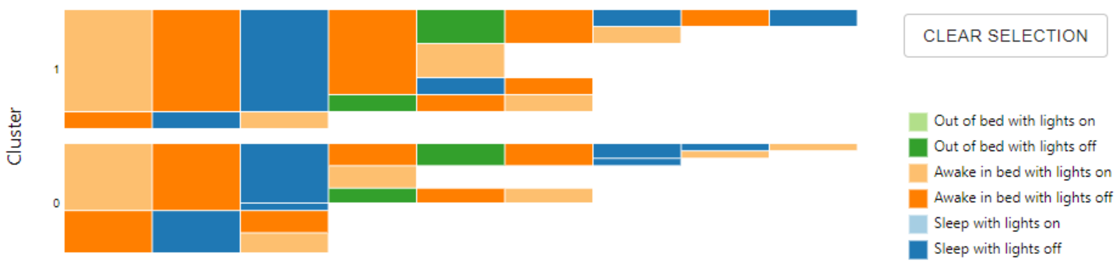

Figure 7.16: The icicle component providing an aggregated overview of the event sequences of two clusters.

The icicle component does not provide information about the start and end times and duration of the events in the event sequences of the clusters. These details can be retrieved from the timeline component. It shows a detailed overview of the event sequences to explore and compare the temporal aspects of the event sequences between the clusters. Figure 7.17 shows how the same event sequence data of Figure 7.16 is depicted in the timeline component. The event sequences of each cluster are placed along the vertical axis accordingly. The timeline shows that the event sequences of cluster 1 belong to nights where subjects went to bed, and usually woke up, earlier than the event sequences that belong to cluster 0.

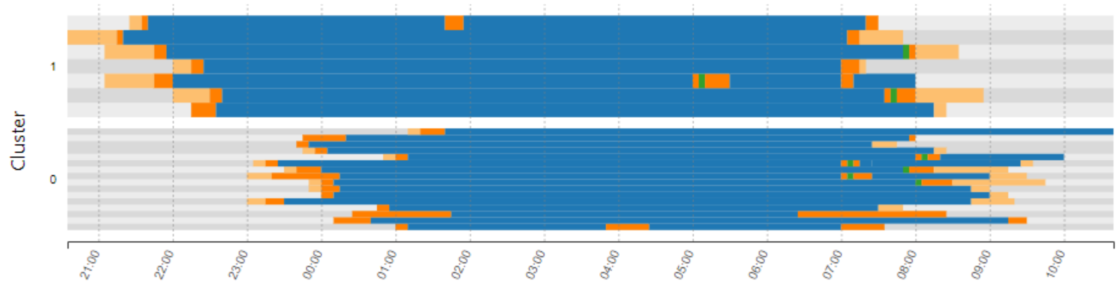

Figure 7.17: The timeline component providing a detailed overview of the event sequences of two clusters.

7.8 Interactions

All components of SleepVis are linked to facilitate selection on all aspects of the data, supporting requirement **R6**. For example, brushing in the correlation exploration component provides the ability to select a certain range of derived and self-reported sleep parameters. In the parallel coordinates plot, brushing can be applied to multiple axes in order to further specify the selection of interest. To increase the focus on the selected lines, the lines that are not selected will be moved to the background by changing the color to grey and reducing the opacity and thickness of the line, shown in Figure 7.4. The other components update accordingly such that the other aspects of the data in the selection can be analysed in more detail in the other components.

Basic interactions such as zooming and panning in the map component help the users to focus on the sleep parameters in a certain area of interest. Additionally, brushing is supported to select an area of interest. A selected area is indicated by a blue transparent rectangle as shown in Figure 7.6. All other components are updated accordingly such that the users can analyse the derived and self-reported sleep parameters and sleep event sequences of the selection in more detail in the other views to better understand the relationship between them, addressing task **T3**.

Furthermore, we use tooltips when hovering over the data in multiple components to provide details. For example, hovering over a day in the calendar component generates a tooltip with the date and average value of the selected sleep parameter of that day.

7.9 Implementation

SleepVis is implemented as an interactive web application for which ReactJS v17.0.1 [16] and D3.js v6 [6] were used. In order to create the map component, we used Leaflet v1.7.1 [2]. The data was pre-processed using Python v3.7.

Chapter 8

Evaluation

In this chapter we describe how we evaluated SleepVis. First, we describe two uses cases to demonstrate the usefulness of the system to achieve the tasks **T1**-**T3**. This is followed by a user study to understand how the users interact with the system to achieve a number of assignments that are inspired by the tasks.

8.1 Use case 1: understanding a good night's sleep

This use case focuses on understanding the relation between multiple derived and self-reported sleep parameters, addressing task **T1** and **T3**. We want to identify what aspects of sleep result in subjects reporting a high (> 0.8) sleep quality and well rested value. We use the parallel coordinates plot in the correlation exploration component to explore the relation between multiple sleep parameters. We use brushing on the sleep quality and well rested axes to select all values higher than 0.8, as shown in Figure 8.1. The selection reveals that high values for sleep quality and well rested are related to low wake after sleep onset values, which means that the subjects were not awake in bed for a long period of time after the first time falling asleep. These values also correspond to low sleep onset latency values, which indicates that it took not long for the subjects to fall asleep. Furthermore, the total sleep time differs between the 300 and 700 minutes and does therefore not necessarily provide a clear indication of what a good night's sleep is and can differ between subjects.

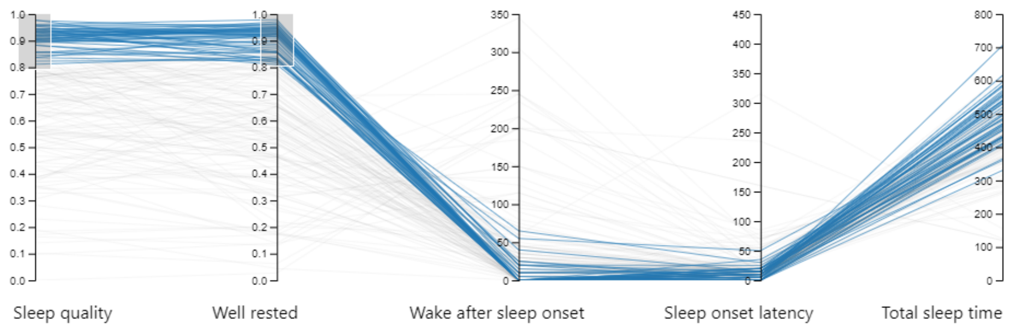

Figure 8.1: The parallel coordinates plot illustrating the relation between sleep quality, well rested, wake after sleep onset, sleep onset latency and total time awake in bed.

We then focus on the cases where subjects report high wake after sleep onset values. We use the parallel coordinates plot to select the nights where subjects were awake for more than 100 minutes after the first time sleep. We are interested in what areas these values are reported. The map component shows that most of the nights with high wake after sleep onset values are reported in the north of Eindhoven, as shown in Figure 8.2. This might be an indication of sleep disruption to external factors such as air traffic or nocturnal noise in the area.

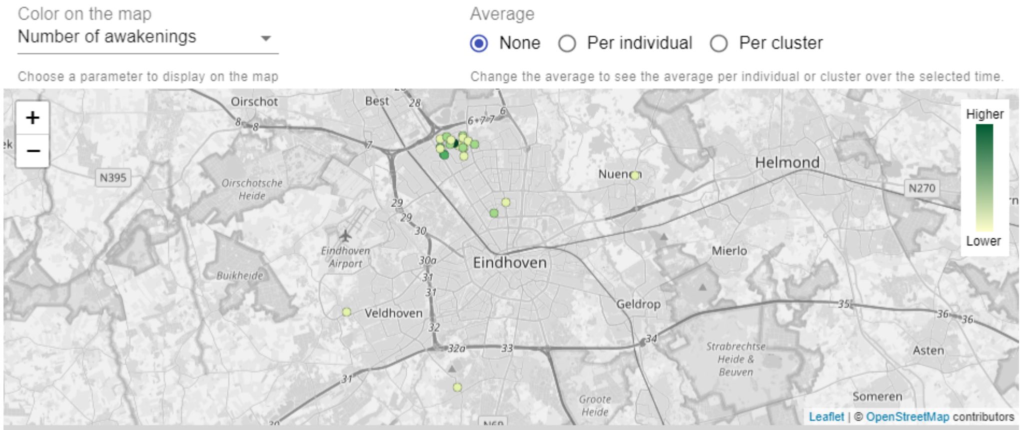

Figure 8.2: The map component showing the location where nights are reported with high wake after sleep onset values.

8.2 Use case 2: find unusual sleep event sequences

The second use case we present addresses tasks **T2** and **T3**. Task **T2** focuses on understanding the behaviour of the sleep event sequence data with respect to location and time. The aim of task **T3** is to understand the relation between sleep parameters and the sleep event sequences.

In this use case we want to know whether there are any unusual patterns in the sleep event sequence data. It is inspired by the insights found by the participants of the user study while exploring the sleep event sequence data. We start the exploration with all subjects and days selected in the year 2021. We use the cluster function to identify deviating sleep patterns in the event sequence data. We expect that the clustering algorithm groups event sequences that are most similar to each other in terms of type and order of events and difference in time. Since we have only a limited amount of data available, we expect the algorithm to create new clusters of event sequence that are extremely different from most of the event sequences. The result shows multiple clusters where one cluster consists of sleep event sequences that only contain *awake in bed with the lights off* and one *awake in bed with the lights on* events. The result of the clustering function are shown in the icicle component in Figure 8.3. Nights where a subject is awake in bed during the complete night is very unusual and requires further exploration.

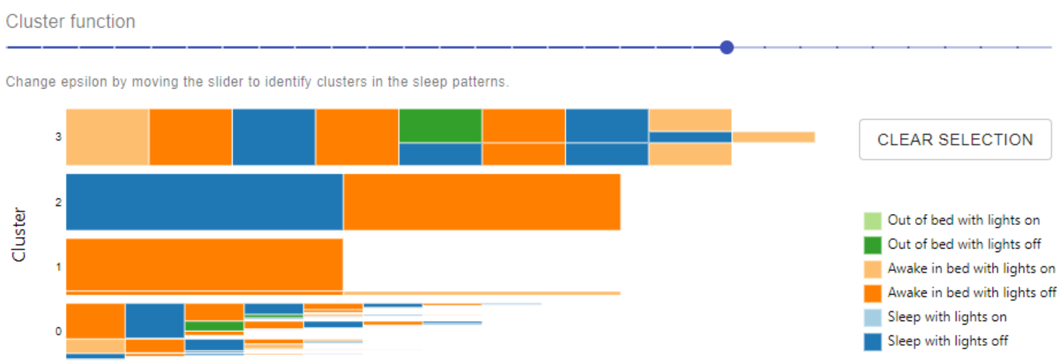

Figure 8.3: Clustering the event sequence data returns one cluster, *Cluster 1*, with unusual sleep patterns.

Next, we select the cluster to analyse the corresponding *sleep quality* and *well rested* values in the scatterplot in the correlation exploration component. It appears that all of the nights, except for one, correspond to a high *sleep quality* and high *well rested* value. It is very unusual that subjects that are awake in bed the complete night would feel rested and thought the quality of the sleep was good. Hence, it is assumed that the sleep periods with the high values were not filled in correctly. We select these sleep periods in the scatterplot for further exploration in the other views.

In the timeline component, we select the *night view* and *group by client* to identify whether the sleep periods belong to one or multiple subjects. The timeline shows only one swimlane, which means that all these night belong to a single subject, shown in Figure 8.4. The tooltip shows that these correspond to subject ID 2.

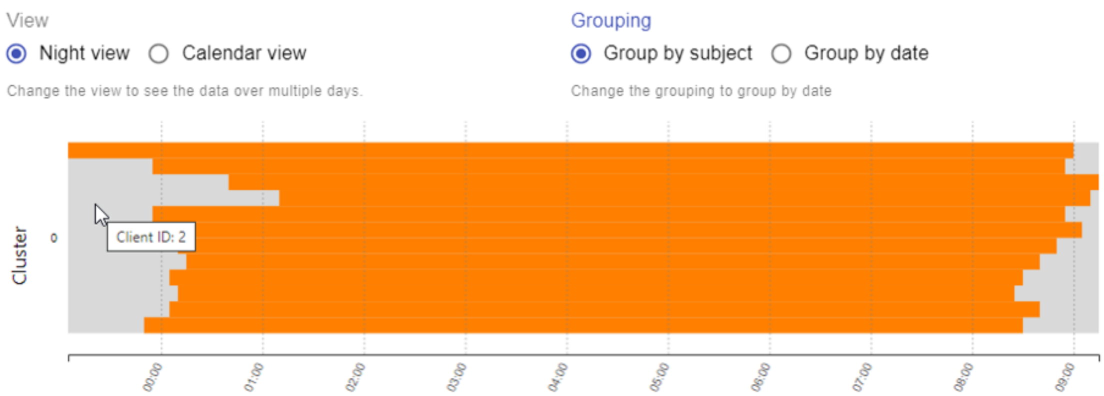

Figure 8.4: Selection of the deviating event sequences shows that they all belong to one subject with subject ID 2.

We clear all the selections, and select only subject ID 2 in the subject selection component. It seems that this subject only reported to be *awake in bed with the lights off*. Hence, we can assume that the subject did not fill in the sleep diary correctly and therefore will not take this subject into account when further analysing the data.

The clustering algorithm is able to detect sleep event sequences that differ much from the majority of event sequences with respect to occurrence and ordering of sleep events. However, most of the sleep event sequences are still grouped together in cluster 0, as shown in Figure 8.3. In order to explore the patterns in this cluster, we select cluster 0 and can perform the cluster analysis on this subset of the data. This reveals another interesting cluster that contains nights of a subject that lives close to the airport, shown in Figure 8.5. The figure shows that the subject often goes out of bed and lies awake in bed between 01:00 and 02:00. The subject also often wakes up at 06:00 and lies awake in bed for some time. These patterns could be an indication of sleep disruption due to external factors such as air traffic.

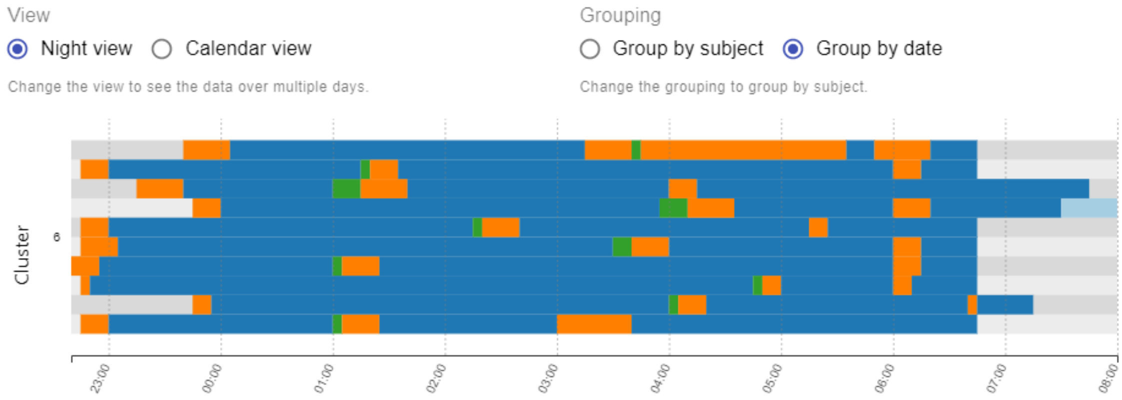

Figure 8.5: The timeline component showing a cluster of event sequences that is revealed by performing a cluster analysis on the event sequences that belong to cluster 0.

8.3 User study

A user study was performed to see whether the users can use SleepVis to complete tasks **T1** - **T3**. First, we describe the method and participants who were involved in the user study. This is followed by a discussion of the results.

8.3.1 Method and participants

The user study consists of two parts, a number of assignments which is followed by a small questionnaire. During the first part of the user study, eleven assignments have to be performed that are based on tasks **T1** - **T3** that we defined in Section 2.3. For example, one of the assignments is to explore the correlation between the self-reported sleep quality and how rested people feel when waking up, addressing task **T1**. The thinking-aloud method [24] is used during this first part to get a clear insight into how the users use SleepVis. The second part of the user study consists of a small questionnaire, designed in Google Forms. The questions use a 7-point Likert scale [27] to indicate the usefulness of each individual component. The last two questions are open questions regarding what features are most useful and what could be improved.

Before actual execution of the user study, a test run was performed to check whether the assignments and questions could be performed and answered within one hour. The subject who participated in the test run had no previous experience with SleepVis and no domain knowledge. The test run took less than one hour so we decided that the number of assignments was appropriate. During the test run we noticed that the questions in which the user had to rate each individual component were too general. Hence, we decided to make them more specific and add a couple of questions about the usefulness of the interaction techniques. For example, *'I found the map component useful to select an area of interest'*. The closed questions from the questionnaire are shown in Table 8.1. We also decided to add one open question about whether the user found any interesting insight during exploration of the data. A complete overview of the assignments and questions can be found in Appendix B.

The user study was performed with two participants. Both the participants are experienced sleep researchers and are familiar with the data and involved in the design of SleepVis. We first provided the participants with a short demonstration of the tool to explain each component and functionality of the system. This was followed by the assignments and questionnaire. The dataset that was used in the user study consists of sleep diary data that was collected with a mobile sleep diary application between the 1st of January 2021 and the 10th of August 2021. In the next section, we discuss the results of the user study.

8.3.2 Results

The participants were able to perform all assignments correctly and answer all the questions within the given time of one hour. It is important to note that the dataset that we used in the evaluation is limited. The dataset is still growing and it is therefore difficult to perform an extensive evaluation.

Assignments

The thinking-aloud method provided several useful insights in how SleepVisis used. Most of the assignments could quickly be performed using the correct component and interaction techniques, such as the first two assignments where the participants were asked to look at the relation between two and three sleep parameters respectively.

Other assignments required more time and effort. Especially the assignments where the participants had to combine multiple views and selections. One example is the assignment where the users were asked to compare a specific sleep pattern on two consecutive days. It was not clear in what order the steps should be performed. Whether they first had to select the event sequence of interest in the icicle component, or look in the timeline component or select the days in the calendar. Hence, the multiple selection options caused confusion.

Another observation is that the participants had the tendency to select the sleep parameter of interest in the correlation exploration component, even if the assignment included comparison of a sleep parameter between days or locations which could be solved using the calendar component or map component respectively. However, the participants indicated that it became more clear where to select the parameter of interest after some time.

Questionnaire

Table 8.1 shows the answers to the closed questions of the questionnaire. The results show that components related to the visualisation of the derived and self-reported sleep parameters were in general evaluated more useful than the components that focused on displaying the sleep event sequence data. One of the participants scored the average option a four out of seven with regard to the usefulness. The participant explained that it is not clear that the average function focuses on both the correlation exploration and map component. It would be more intuitive to have a separate average function for the plot and map component.

The participants indicated that they liked the scatterplot and parallel coordinates plot most. The selection of the sleep parameters of interest and brushing worked very intuitively.

When we asked the participants what could be improved, one response was that in the correlation exploration component it was not possible to see whether the circles or lines belonged to one or multiple subjects. They provided the suggestion to add the option to change the color coding of the circles and lines in the view such that they indicate the subject, instead of the cluster.

The participants also indicated that it was difficult to select all subjects except for one in the subject selection drop-down menu. Therefore, the participant suggested to add the option to select all subject IDs. Other features that would be useful are the option to upload a dataset and to export the selection such that it can be used for statistical analyses.

We also asked the participants whether they found any interesting insights while exploring the data. One of the answers was that it was interesting to see that the self-reported sleep quality was lower in the area close to Eindhoven Airport than in the center of Eindhoven. The other participant responded: *"Although the cluster function was a bit difficult to understand, I think it could be useful to e.g. cluster types of insomnia patients by sleep pattern. Perhaps with this tool different types of insomnia patients could be identified with different types of issues during the night (e.g. a cluster with heavily fragmented sleep vs a cluster with one long extended period of wakefulness)."* The clustering algorithm we used should be able to detect these differences in sleep patterns if the differences are large with respect to time and ordering and occurrence of sleep events, as demonstrated in the second use case.

Overall, the users rated SleepVis positively. *"The tool allows exploration of the sleep diary data on many different levels."*

Question	1 (Strongly disagree)	2	3	4	5	6	7 (Strongly agree)
I found the option to select a subject of interest useful.						X	X
I found the scatterplot useful to explore the relationship of two sleep parameters.							X X
I found the parallel coordinates plot useful to explore the relationship of multiple sleep parameters.						X X	
I found the option to change the order of the axes of the scatterplot and parallel coordinates plot useful to observe the relationship between sleep parameters.						X	X
I found the brush function of the scatterplot and parallel coordinates plot useful to select a range of interest of certain sleep parameters.						X X	
I found the map useful to observe the distribution of sleep parameters over location.						X X	
I found the map useful to select an area of interest.						X X	
I found the average option useful to observe the average value of sleep parameters of an individual or cluster.						X	X
I found the calendar useful to observe differences in sleep parameters between weekdays and weekends, months and seasons.				X		X	
I found the calendar useful to select days of interest.						X X	
I found the icicle plot useful to get an overview of all sleep patterns in the data.					X X		
I found the icicle plot useful to select sleep patterns of interest.					X X		
I found the timeline useful to get a detailed overview of the sleep patterns in the data.					X	X	
I found the timeline useful to select a time range of interest to observe local sleep parameters.					X	X	
I found the cluster function useful to identify different sleep patterns.					X	X	

Table 8.1: Results of the closed questions of the questionnaire that was used in the user study.

8.4 Final development of SleepVis

In the final version of SleepVis we implemented some of the improvements that were suggested by the participants of the user study. The participants indicated that it was confusing that the average function focused on both the correlation exploration and map component. Therefore, we created a separate function for the correlation exploration component and the map component to average the data, as shown in Figure 8.6.

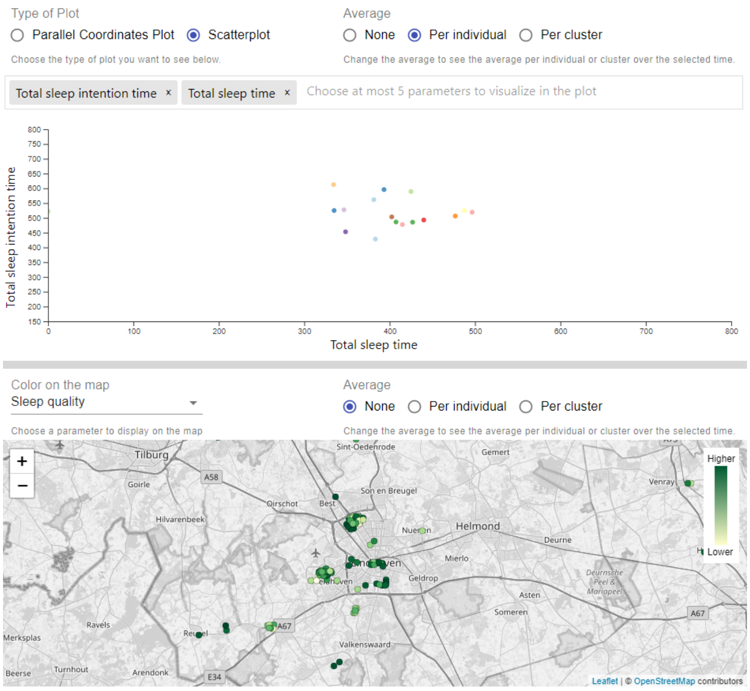

Figure 8.6: The final design of SleepVis which has a separate average function for the correlation exploration component and map component.

Another remark of the participants of the user study was that it was difficult to select all subject IDs. This required selecting each subject individually in the drop-down menu. Hence, we implemented the option to select all and remove all subject IDs as shown in Figure 8.7.

Figure 8.7: The final design of SleepVis has the option to select all subject IDs.

Chapter 9

Conclusions and future work

In this chapter, we first present the main conclusions. This is followed by a discussion about potential future work in the domain of visualization with regards to analysis of sleep diary data.

9.1 Conclusions

The goal of this project was to build a visual analytics system to analyse data from sleep diaries to help sleep experts understand the behaviour of derived and self-reported sleep parameters and sleep event sequences to formulate new questions and hypotheses. In order to achieve this goal, we analysed the domain situation and data and the tasks **T1-T3**. Based on the data and tasks, we presented related literature which served as the inspiration of our design choices.

The visual encodings were designed for exploration and identification of trends and outliers. The interactive solutions supported exploration of the geographical aspect of the data and analysis at multiple temporal levels. We also provided a way to summarize the event sequence data and to group similar event sequences. The clustering algorithm was able to detect especially large differences between event sequences with respect to time and ordering and occurrence of events in a limited dataset. This helped the users to find trends and outliers in the raw sleep event sequence data.

We provided two use cases to illustrate how SleepVis can be used to achieve tasks **T1-T3**. We presented how multiple components and selections can be combined to find patterns and outliers in the multivariate event sequence data. Additionally, a user study was performed with two sleep experts to evaluate the usefulness of SleepVisfor exploration of sleep diary data. The assignments covered each individual component and related tasks. The usefulness of the visualizations and interactions techniques were assessed by the questionnaire. Based on the comments of the participants we have made some small improvements in the final design of SleepVis, such as the option to select all and remove all subjects from the selection. Overall, the usefulness of the elements received a high rating and the impressions of SleepVis were positive. SleepVis shows promising results, however, more evaluation is needed. Due to time constraints, the system has only been evaluated by two users who were involved in the design of SleepVis. Furthermore, the dataset was limited in size which made it difficult to perform an extensive evaluation. In order to better evaluate the system, a more elaborate user study should be performed with a larger dataset.

9.2 Future work

In this project, we have identified several potential extensions to improve the system:

- The similarity measure that we used in this project to define similarity between event sequences could be improved. The metric treats all types of sleep events equally. However, some sleep events are closer to one type of sleep event than to another type of sleep event.

For example, the difference between *sleep with the lights off* and *out of bed with the lights off* is bigger than the difference between *sleep with the lights off* and *sleep with the lights on*. This is currently not taken into account by the measure. Therefore, there is still room for improvement with regard to defining the similarity between event sequences.

- One of the participants of the user study indicated it would be useful to have the option to upload other datasets and export the selection. With this function, the users will be able to use the system for the exploration of external datasets.
- Currently, the icicle plot is useful to get a summarized overview of all event sequences. However, focus and context and the option to zoom would enhance the exploration of the event sequence data. Outliers in the event sequences are currently hard to identify since they are represented by thin bars. The option to zoom on the horizontal and vertical axis would allow the users to identify the deviating event sequences more easily.
- The raw dataset contains notes that the subjects added to the sleep periods. These notes could be helpful in better understanding the behaviour of the derived and self-reported sleep parameters and sleep event sequences since they can give insights into why people did not sleep well. Natural language processing techniques could help with this.
- As discussed in Chapter 1, the sleep diary data is part of a study on the relation between air traffic and sleep. The next step would be to incorporate the air traffic into the visualizations to explore whether there is a direct relation between the sleep and air traffic. This project focused on extensive analysis of the sleep diary data which makes SleepVis a useful starting point for a solution to gain insight into the relation between sleep and air traffic.

Bibliography

[1] United States Naval Academy. Approximate metric equivalents for degrees, minutes, and seconds. 21

[2] V. Agafonkin. Leaflet: An open-source javascript library for mobile-friendly interactive maps. 41

[3] M. van Ballengooij, A.M. van Gestel, and M. Scholtes. Beleving leefomgeving rondom vliegveld Eindhoven, 2018. 1

[4] J. Blake. Development of an online sleep diary for physician and patient use. *Knowledge Management & E-Learning: An International Journal*, page 188–202, 2010. 1, 12

[5] J. Blake and D. Kerr. Sleep disorder diagnosis: the design and implications of online tools. *Decision Analytics*, 1:1–18, 2014. ix, 12

[6] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. 41

[7] C. Brewer, M. Harrower, and The Pennsylvania State University. Colorbrewer 2.0: color advice for cartography. 32, 34, 35

[8] B. Cappers and J.J. van Wijk. Exploring multivariate event sequences using rules, aggregations, and selections. *IEEE Transactions on Visualization and Computer Graphics*, 24:532–541, 2018. 15

[9] M. Ester, H.-P Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996. 29

[10] H. Garcia Caballero, M. Westenberg, B. Gebre, and J. van Wijk. V-awake: a visual analytics approach for correcting sleep predictions from deep learning models. *Computer Graphics Forum*, 38(3):1–12, March 2019. 21st Eurographics/IEEE VGTC Conference on Visualization ; Conference date: 03-06-2019 Through 07-06-2019. ix, 12, 13

[11] M.A. Grandner, N.J. Jackson, V.M. Pak, and P.R. Gehrman. Sleep disturbance is associated with cardiovascular and metabolic disorders. *Journal of Sleep Research*, 21(4):427–433, August 2012. 1

[12] H. Guo, Z. Wang, B. Yu, H. Zhao, and X. Yuan. Tripvista: Triple perspective visual trajectory analytics and its application on microscopic traffic data at a road intersection. pages 163–170, 03 2011. 16, 31

[13] M. Harrower and C.A. Brewer. *ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps*, chapter 3.8, pages 261–268. John Wiley and Sons, Ltd, 2003. 32

[14] J. Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. 1997. 25

[15] J. Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2:283–304, 2004. 25

[16] Facebook Inc. React: A javascript library for building user interfaces. 41

[17] S.J. Kim, S.K. Chai, K.W. Lee, J. Park, K. Min, H.G. Kil, C. Lee, and K.J. Lee. Exposure–response relationship between aircraft noise and sleep quality: A community-based cross-sectional study. *Osong Public Health and Research Perspectives*, 5(2):108 – 114, 2014. 1

[18] B. Koseoglu, E. Kaya, S. Balcisoy, and B. Bozkaya. St sequence miner: visualization and mining of spatio-temporal event sequences. *The Visual Computer*, pages 1 – 13, 2020. 15, 33

[19] J. Kruskal and J. Landwehr. Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37:162–168, 1983. 14

[20] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013. 14, 38

[21] S. Morrell, R. Taylor, and D. Lyle. A review of health effects of aircraft noise. *Australian and New Zealand Journal of Public Health*, 21(2):221–236. 1

[22] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015. 2, 5, 8, 31

[23] A. Muzet. Environmental noise, sleep and health. *Sleep Medicine Reviews*, 11(2):135 – 142, 2007. 1

[24] J. Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994. 46

[25] H. Nikamalfard, H. Zheng, H. Wang, M. Mulvenna, P. McCullagh, and P. Jeffers. A sleep pattern analysis and visualization system to support people with early dementia. In *2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*, pages 510–513, 2011. ix, 13

[26] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. *LifeLines: Visualizing Personal Histories*, page 221–227. Association for Computing Machinery, New York, NY, USA, 1996. 14

[27] J. Robinson. *Likert Scale*, pages 3620–3621. Springer Netherlands, Dordrecht, 2014. 46

[28] J. J. van Wijk and E. R. van Selow. Cluster and calendar based visualization of time series data. In *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 4–9, 1999. 34

[29] S. Wallace, H. Guo, and D. Sasson. Visualizing self-tracked mobile sensor and self-reflection data to help sleep clinicians infer patterns. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 2194–2200, New York, NY, USA, 2017. Association for Computing Machinery. ix, 11

[30] K. Wongsuphasawat, J. Gómez, C. Plaisant, T. D. Wang, Meirav Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011. 2

[31] K. Wongsuphasawat and B. Shneiderman. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 27–34, 2009. 26

[32] R. Yi, M. Enayati, J. Keller, M. Popescu, and M. Skubic. Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor. 07 2019. ix, 12

Appendix A

Sleep parameter calculation

List of symbols

Notations	Definitions
C	Set of categories such that $C = \{$*Sleep with lights off, Sleep with lights on, Awake in bed with lights off, Awake in bed with lights on, Out of bed with lights off, Out of bed with lights on*$\}$
S	Set of all events sequences s in the data
s	Event sequence s such that $s = \{(t,c) \mid t \in Time \text{ and } c \in C\}$
t_{e_0}	Start time of event e
t_{e_1}	End time of event e
t_{r_0}	Start time of time range r
t_{r_1}	End time of time range r

Equations

Terminal wake time

$$t_{\mathit{final\ awakening}} - t_{\mathit{lights\ on}} \tag{A.1}$$

Total bedtime

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.2}$$

where $A = \forall e[e \in s | e = \mathit{Awake\ in\ bed\ with\ lights\ on} \vee e = \mathit{Awake\ in\ bed\ with\ lights\ off} \vee e = \mathit{Sleep\ with\ lights\ on} \vee e = \mathit{Sleep\ with\ lights\ off}]$ and $s \in S$

Total intended wake in bed time

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.3}$$

where $A = \forall e[e \in s | e = \mathit{Awake\ in\ bed\ with\ lights\ on}]$ and $s \in S$

Total sleep intention time

$$t_{\mathit{lights\ on}} - t_{\mathit{lights\ off}} \tag{A.4}$$

Total sleep time

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.5}$$

where $A = \forall e[e \in s | e = \mathit{Sleep\ with\ lights\ on} \vee e = \mathit{Sleep\ with\ lights\ off}]$ and $s \in S$

Total time awake in bed

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.6}$$

where $A = \forall e[e \in s | e = \mathit{Awake\ in\ bed\ with\ lights\ on} \vee e = \mathit{Awake\ in\ bed\ with\ lights\ off}]$ and $s \in S$

Total time out of bed

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.7}$$

where $A = \forall e[e \in s | e = \mathit{Out\ of\ bed\ with\ lights\ off}]$ and $s \in S$

Wake after sleep onset

$$\sum_{e \in A} t_{e_1} - t_{e_0} \tag{A.8}$$

where $A = \forall e[e \in s | e = \textit{Awake in bed with lights off} \lor e = \textit{Out of bed with lights off}]$ and t_{e_1} and $t_{e_0} > t_{\textit{first sleep}_1}$ and $s \in S$

Sleep onset latency

$$t_{\textit{first sleep}_0} - t_{\textit{lights off}} \tag{A.9}$$

Local bedtime

$$\sum_{e \in A} min(t_{e_1}, t_{r_1}) - max(t_{e_0}, t_{r_0}) \tag{A.10}$$

where $A = \forall e[e \in s | e = \textit{Awake in bed with lights on} \lor e = \textit{Awake in bed with lights off} \lor e = \textit{Sleep with lights on} \lor e = \textit{Sleep with lights off}]$ and $s \in S$

Local sleep efficiency

$$\textit{Local sleep time}/\textit{Local bedtime} * 100\% \tag{A.11}$$

Local sleep intention time

$$min(t_{\textit{lights on}}, t_{r_1}) - max(t_{\textit{lights off}}, t_{r_0}) \tag{A.12}$$

Local sleep time

$$\sum_{e \in A} min(t_{e_1}, t_{r_1}) - max(t_{e_0}, t_{r_0}) \tag{A.13}$$

where $A = \forall e[e \in s | e = \textit{Sleep with lights on} \lor e = \textit{Sleep with lights off}]$ and $s \in S$

Local time awake in bed

$$\sum_{e \in A} min(t_{e_1}, t_{r_1}) - max(t_{e_0}, t_{r_0}) \tag{A.14}$$

where $A = \forall e[e \in s | e = \textit{Awake in bed with lights on} \lor e = \textit{Awake in bed with lights off}]$ and $s \in S$

Local time out of bed

$$\sum_{e \in A} min(t_{e_1}, t_{r_1}) - max(t_{e_0}, t_{r_0}) \tag{A.15}$$

where $A = \forall e[e \in s | e = \textit{Out of bed with lights off}]$ and $s \in S$

Total awakenings

$$\sum_{e \in A} 1 \tag{A.16}$$

where $A = \forall e[e \in s | e = \textit{Awake in bed with lights off}]$ and $s \in S$

Total out of bed

$$\sum_{e \in A} 1 \tag{A.17}$$

where $A = \forall e[e \in s | e = \textit{Out of bed with lights off}]$ and $s \in S$

Local sleep efficiency

$$\textit{Total sleep time}/\textit{Total bedtime} * 100\% \tag{A.18}$$

Total sleep episodes

$$\sum_{e \in A} 1 \tag{A.19}$$

where $A = \forall e[e \in s | e = \textit{Sleep with lights off} \vee e = \textit{Sleep with lights on}]$ and $s \in S$

Appendix B

User evaluation SleepVis

User evaluation SleepVis

*Vereist

Assignments

1. Is there a correlation between reported sleep quality and how rested people feel when waking up?

2. Are there any people that reported a sleep quality between 0.6 and 0.8, went out of bed two times and slept for more than 8.5 hours (510 minutes)?

3. What is the average number of times out of bed and the average number of awakenings of client 4?

4. Is the sleep efficiency on July 29 2021 higher in Woensel Noord than in the center of Eindhoven?

5. Is there a clear difference between the times that client 15 turns on the lights during weekdays and weekends?

6. Are there any subjects that are awake in bed with the lights turned off the whole night? If yes, what is the client ID of the subject(s)?

7. What is the number of nights that people living in Waalre don't go out of bed at night?

8. Have more people been awake in bed the whole night on the 6th of August than on the 7th of August?

9. Are the average local sleep efficiency and local sleep intention time of client 15 higher between 01:00-02:00 or 05:00-06:00?

10. Are there any people that had the following sleep pattern: 'Awake in bed with the lights on - Awake in bed with the lights off - Sleep - Awake in bed with the lights on', and reported a sleep quality lower than 0.7? If yes, what is the client ID of the subject(s) and on what day?

11. How many clusters are defined with an epsilon of 1.7? How many people are in cluster 1? How many different days are in cluster 2?

Questions

12. I found the option to select a subject of interest useful. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

13. I found the scatterplot useful to explore the relationship of two sleep parameters. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

14. I found the parallel coordinates plot useful to explore the relationship of multiple sleep parameters. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

15. I found the option to change the order of the axes of the scatterplot and parallel coordinates plot useful to observe the relationship between sleep parameters. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

16. I found the brush function of the scatterplot and parallel coordinates plot useful to select a range of interest of certain sleep parameters. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree	◯	◯	◯	◯	◯	◯	◯	Strongly agree

17. I found the map useful to observe the distribution of sleep parameters over location. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree	◯	◯	◯	◯	◯	◯	◯	Strongly agree

18. I found the map useful to select an area of interest. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree	◯	◯	◯	◯	◯	◯	◯	Strongly agree

19. I found the average option useful to observe the average value of sleep parameters of an individual or cluster. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree	◯	◯	◯	◯	◯	◯	◯	Strongly agree

20. I found the calendar useful to observe differences in sleep parameters between weekdays and weekends, months and seasons. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

21. I found the calendar useful to select days of interest. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

22. I found the icicle plot useful to get an overview of all sleep patterns in the data. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

23. I found the icicle plot useful to select sleep patterns of interest. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

24. I found the timeline useful to get a detailed overview of the sleep patterns in the data. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

25. I found the timeline useful to select a time range of interest to observe local sleep parameters. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

26. I found the cluster function useful to identify different sleep patterns. *

Markeer slechts één ovaal.

	1	2	3	4	5	6	7	
Strongly disagree								Strongly agree

27. What features of SleepVis do you like most? *

28. What do you think could be improved or is missing? *

29. Did you find any interesting insights during exploration of the data? If yes, what insight did you find? *