# Stress data visualization

Semikina, S.A.

*Award date:*
2014

Technische Universiteit Eindhoven
University of Technology

Department of Mathematics and Computer Science

# Stress Data Visualization

*Master Thesis*

Sofia Semikina

Supervisors:

Natalia Sidorova
Rafal Kocielnik
Michel Westenberg

Eindhoven, March 2014

# Abstract


Stress related research is becoming nowadays increasingly popular. Special attention is devoted to stress at work. Recent studies provide proofs of dangerous influence that stress has on human health. It leads to the appearance of a new important part of self-healthcare: stress monitoring.

Existing ways of stress monitoring include various ways of data collection, visualization and representation to the user. The focus of the current work is on the second part: stress data visualization techniques.

One of the problems of existing visualization techniques for stress level is isolation of stress data from the real life of a person, experiencing emotional arousal. It is not always easy to associate some specific feature showed in the visualization with an event taking place at that particular moment of time, since it requires good memory from the user. However, in context of stress experienced at work, this should be possible, because work time usually has some structure and events are often scheduled for specific time moments. In this work, we try to put together two sources of important information: stress level data of a user and a personal calendar, containing information about professional and personal life events.

After taking a look at existing techniques for visualization of time-series data, we discuss how some of these techniques can be adapted for representation of stress level of a person. We propose a number of new ideas for stress data visualization, and possible ways of combining it with calendar information. We discuss different alternatives of stress data aggregation by time (per minute, half a day, and one day) and by event types, embedding of the visualizations into the calendar views (arrangement by day, week, and month), variations on number of users shown on a visualization. Several alternatives are validated in a small user study. The validation results prove that there is a visualization technique prefered by most users. To gain a better understanding of user view on the month visualization, another study with a small group of participants is performed. The results of the study show differences in user perception depending on the visualization technique, even if the same piece of data is visualized; some similarities in men and women perception of visualizations; existence of different groups of users having the similar view on the questions and factors defining the choice of each group.

To prepare the use of validated visualization ideas in real applications, we investigate possible ways of storage and management of sensor stress level data. Considered alternatives are evaluated in terms of performance, ease of access and processing data, maintainability, and memory use, and conclusions about the most appropriate way of data organization are made.



**Keywords:** stress, data visualization, time-series data

# Preface

This Master's thesis is the result of my graduation project at the Department of Mathematics and Computer Science. The graduation project is a part of my Master of Computer Science and Engineering study at Eindhoven University of Technology. I would like to express my gratitude to all people who helped me to finish my thesis.

First of all, I would like to thank Natalia Sidorova for her continuous supervision and guidance during this project. I am very grateful for the time she has spent reading and giving feedback on my work, discussing ideas and giving insights. She has always kept me motivated to move forward, being encouraging and positively confident about the result. I also deeply appreciate the help of my graduation tutor Rafal Kocielnik for being available whenever I had questions, his support, kindness, and patience while explaining to me complicated details. Many thanks to Joos Buijs who was very kind to share with me the template of his master's thesis that I used to write this paper. Furthermore, I would like to thank all participants of the user studies conducted for this project for investing their time and effort.

I also owe my success to my former supervisor from Moscow Irina Lomazova. Without her idea and effort on creating the double master program it would hardly become possible for me to reach this point.

My sincere thanks to my friends who encouraged me all this time, especially to Ivan, Sofia and Mira, whose help and knowledge was always very useful and timely. And my biggest and warmest thanks go out to my parents and my grandmother. Without their love, trust in me and support I would never get that far.

# Contents

# Chapter 1

# Introduction

Modern life puts high pressure on a person. Living in a fast-paced world, being always in a hurry, having to deal with various professional or personal problems, making too high demands on themselves make people feeling stressed. Obviously, some stress is inevitable and it can even be helpful, giving extra energy to meet a deadline or a challenge, but continuous high stress level can lead to raised anxiety and various diseases, affecting both physical and emotional health of a person. Thus, there is a need in stress management, which means carefully balancing it, keeping it at a reasonable level and not letting it to reach dangerous limits.

Lately, special attention has been given to the role of stress at work. Indeed, if compared to a worker with a medium stress level, a highly stressed employee is as a rule lower motivated and less effective, and hence is more expensive for a company. Moreover, recent studies reported that mental disorders called by stress, especially depression, are growing reasons for work disability and early retirement. Thus, stress balancing is an urgent problem that is of interest for many companies. That is one of the reasons for stress-concerned research becoming increasingly popular.

## Problem Description

Current approaches to stress monitoring demonstrate a lot of variety and creativity. That concerns both data collecting methods and ways of its representation to the users. The techniques of stress data obtaining and visualizing are described in more detail in section 2.1.

Some of the existing visualizations of the stress data make it possible to see the stress level in dynamics, to notice increases, falls or individual peaks of the stress level, periods of fluctuation or stabilization, etc. Other visualization techniques use aggregated values of the stress level, allowing for pattern discovery and trends definition. All kinds of information that users are able to grasp from the visualizations help them to learn about their stress and understand it better. But in real life, the stress level value at a certain point of time is often a result of life events of the person at that particular moment, or before/after it, e.g., attending a personal/professional meeting, receiving a mail, talking to a colleague, etc. Trying to establish a connection between the stress level and life events often requires much effort from the user, and sometimes it is impossible at all. Existing visualizations barely allow for convenient matching of stress level data to work and personal life events at the instance level.

We aim at the presenting stress level data in relation to everyday life events to enable the recall function of the user. Recall function makes users able to remember what happened at the time moment of the day, when the visualization features a peak, what sequence of actions led to a significant increase in the stress level, or what helped to maintain a reasonable stress level during some day. Becoming able to associate some specific feature of the stress visualization with a definite life event, users will get more insight in their stress, understand better how to manage it, learn from this data how to behave in future in order to balance the stress level.

> ***Problem Statement:*** *Existing visualizations present stress level data in isolation from information about life events, which makes it difficult for the user to understand his/her stress data and learn from it.*

Having the problem formulated, it becomes now possible to define the goal of the project.

## Project Goal

Having connection between the stress level and life events, users will be able to gain new knowledge about their stress.

*The goal of this project is to enable the recall function of the users by putting different sources of information (life events, stress level data) together by developing visualization technique(s) for the stress data and decide about the way that the stress data and life events data should be stored.*

### Visualization Aim

A traditional aim of any visualization is insight. Due to the phenomenal abilities of the human eye, people perceive visual data much easier than the same data in form of numbers; visualization makes it possible to detect structures in images, to notice patterns and trends. It makes the users able to see things they were not aware of, and to define new questions, hypotheses, and models of the visualized data. However, there is a strange paradox in the basic paradigm of visualization. We do not know what information is contained in the data, hence we make pictures to get insight, but if we do not know which specific aspects or features should be visible, we cannot assess if we are successful or not [20]. Thus, first it is necessary to define what should be seen in the visualizations we are building, what kind of information users should be able to extract from the views, what are the questions to be answered by users.

The main goal of the user is to decide how to behave, which actions to take in order to balance the stress level. Thus, the user should be enabled to get relevant information from the visualizations. Since we aim at enabling the recall function of the user, we want to put several sources of information together. The personal calendar of the user and visualization of his/her stress level data obtained from the sensors visualized together should be able to bring the answers to such questions as: which events of the day/week are the most/least stressful, what time of day is usually the most stressful/calm, at which moments the stress level changes steeply/gradually, which day demonstrates the biggest/smallest difference between the highest and the lowest stress level, etc.

Based on the questions to be answered, existing visualization techniques should be analyzed and the most appropriate one(s) should be chosen for visualization implementation.

## Approach

The process of research is divided into several steps.

- First, to become aware of the works in the field of visualization and, in particular, stress visualization, an overview of recent research is done. The knowledge obtained in this study could give insight in applying examined visualization ideas to the problem domain of stress.
- Second, existing visualization techniques are adjusted to representing stress data. Some of existing techniques give an insight for developing new visualization ideas. Then chosen ideas demonstrating the presence of connection between the stress data visualization and life events, are validated with a small group of participants.
- After that, possible alternatives for the data storage and management are considered and the performance evaluation for each of proposed data storage alternatives is done. Finally, based on the results of performance evaluation, the choice of an appropriate data storage is made.

## Outline

The reminder of this paper is structured as follows. The results of the overview of the previous work in visualization field are presented in chapter 2. Special attention is given to the visualization techniques for time-series data and also to stress data visualization. In chapter 3, some of the ideas described in chapter 2 are elaborated and adjusted to the scope of current project. In chapter 4, two small user studies are described. One of them performes the validation of several visualization techniques, the other is aimed at getting an insight of the way of creating one of the visualization ideas reviewed in chapter 2. Chapter 5 addresses the alternatives of data storage, organization and management for building described visualizations for big amounts of data. In this chapter, possible solutions are proposed and some experiments are performed in order to choose the most convenient one in terms of speed of work and amount of space necessary. Main conclusions of the project and discussion on the directions of further project development are discussed in chapter 6.

# Chapter 2

# Related Work

As described in chapter 1, one of the main goals of this work is to develop some new visualization techniques for stress data. Since the field of time-dependent data visualization has a long history of investigation and a lot of useful visualization techniques already exist, it is wise to use the best of existing ideas and to adapt them to the subject domain. Investigations in the area of stress data visualization are discussed in section 2.1. A summary of the previous research in the field of time-oriented data visualization is provided in section 2.2. Some visualizations tools are mentioned in section 2.3.

## 2.1 Stress Data Visualization

There is a rich body of research in the area of stress, proposing a big variety of stress monitoring methods. Some of them are based on applying questionnaires or carrying out individual/group meeting with psychologists. Questionnaires do not always reflect the actual experiences since they rely mainly on peoples feedback on the event, which is biased due to specific features of the human memory. Meetings with a psychologist can be very effective, but are very costly both in terms of money and time [11].

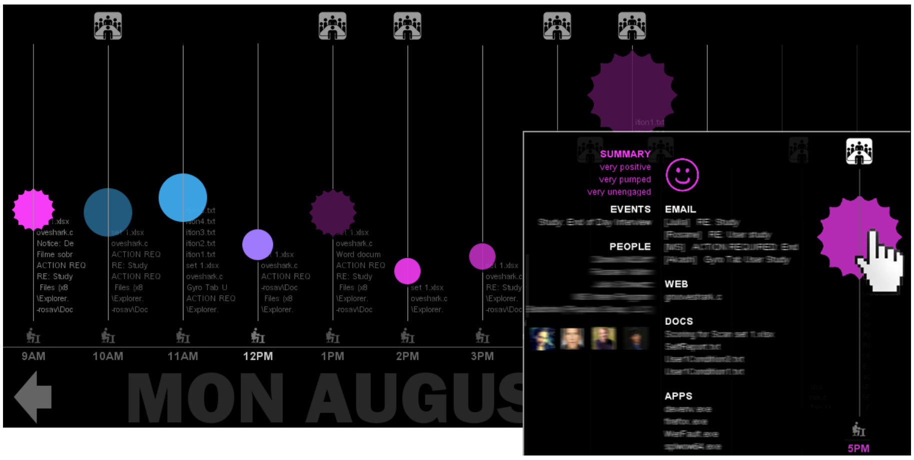



Figure 2.1: An example of stress data visualization (from [13])

A more objective picture of the stress level of a person can be obtained when using modern sensor technology. The overview presented in [17] discusses a number of physical and psychological signals (e.g., heart rate, skin conductance, facial movements, blood pressure, skin temperature, etc.) that can be measured to gain the actual stress data. However, most works in the area of stress detection were performed in lab conditions, where the environment is strictly controlled, the stressful conditions are administered artificially and, as a consequence, do not necessarily conform to the ones experienced in real life. Besides, the measurements in lab settings are short-term and mostly static; almost all the personal, natural context of experiencing stress is removed [11].

Recently several attempts to detect stress level in life-settings were taken. The project "AffectAura" described in [13] aims at emotion detection in real life environment, allowing users to reflect on their emotional states over long periods of time. This project proposes visualizations that include several dimensions representing different sources of information (e.g., visualized sensor data, user's feedback, personal calendar information, activity logs). Figure 2.1 shows an example of visualization technique used in the "AffectAura" project. However, though the presented visualizations enrich the calendar with additional data, they do not reflect dynamical changes during a period of time (e.g. a day, a week), but only operate with aggregated values (the emotional data is aggregated per hour, so a few changes within one day can be observed, but since the emotional state of a person can change more often than once in one hour, current detalization level is quite low).

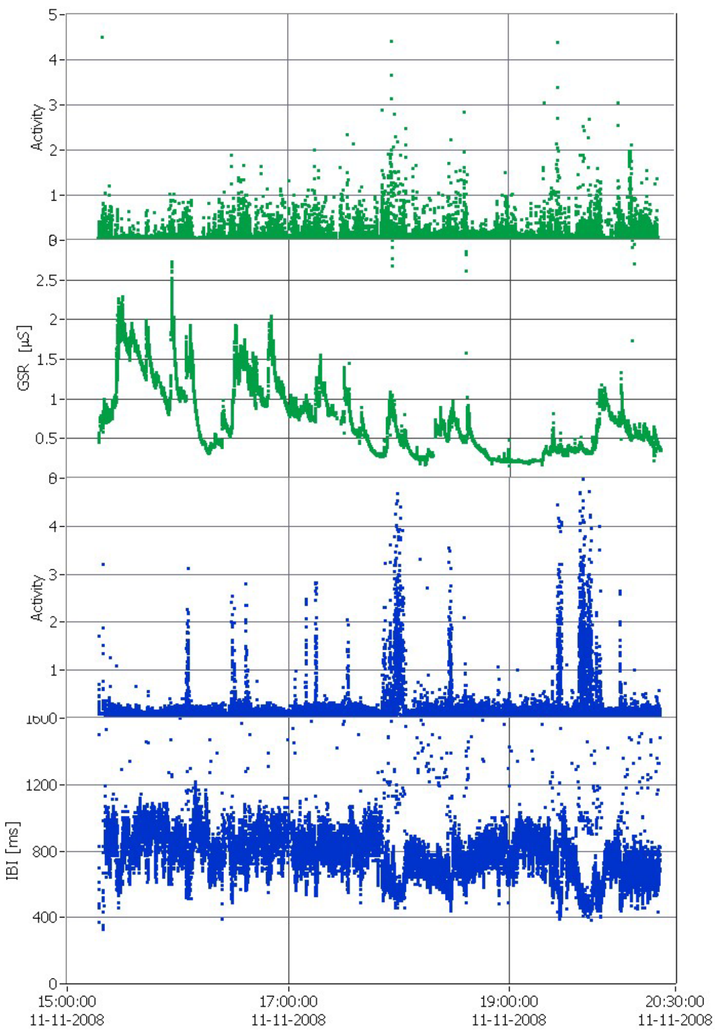



Figure 2.2: An example of stress data visualization (from [22])

The effective and detailed emotion recording in the "AffectAura" project is possible due to the huge amount of sensors used. The technologies applied (webcam, microphone) disturb users' privacy. Since the privacy issue is of high importance for most people, a lot of recent work (e.g., see

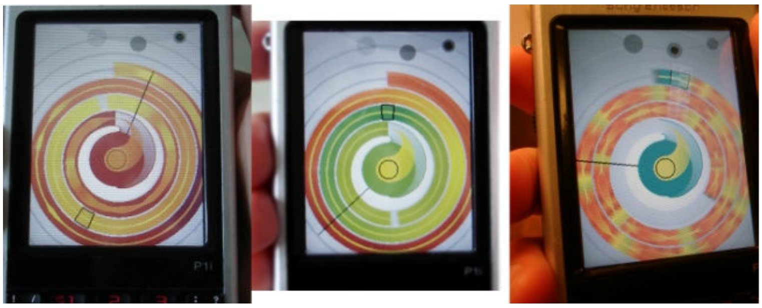


Figure 2.3: An example of stress data visualization (from [16])

[11], [16], [22]) focus on unobtrusive ways of sensor measuring. However, though the measurements in these works are made in real life conditions, visualization applications described in [16], [22] present stress level data in isolation from real life events. Visualizations help users to observe the changes in their stress level, but it often requires much effort (and sometimes is impossible at all) to remember what led to this or that level of stress, which events happened that day, why the stress level grew and fell. Figures 2.2 and 2.3 illustrate the visualization kinds used in the projects [22] and [16] respectively.

A number of existing applications try to solve the problem of the stress data being disconnected from the life events. The goal of the framework presented in [11] is showing stress information derived from sensor measurements in the context of persons activities (information about life events from digital calendars, social media, emails, logs of phone calls, etc.). Presenting the data collected from sensor in relation to digital life information, as demonstrated in figure 2.4, enables the user to discover personal stress and relief patterns. But since the visualized data is aggregated for the chosen period of time, there is still no opportunity to identify individual peaks, spot increases, falls and fluctuations.

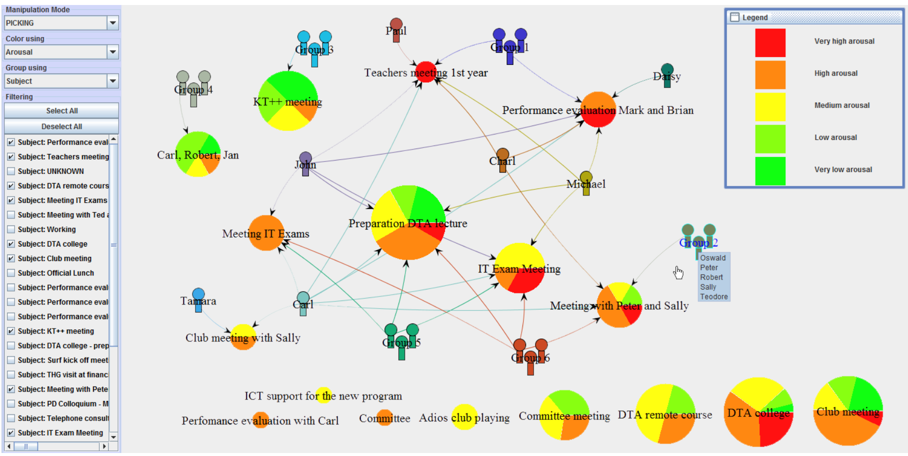



Figure 2.4: An example of stress data visualization (from [11])

## 2.2 Time Series Data Visualization

In this section, an overview of related work on time series data visualization is presented. In section 2.2.1, the criteria for classification of visualization techniques for time series data are discussed. Section 2.2.2 provides some examples on the proposed classification.

### 2.2.1 Classification

The systematic view on techniques for visualizing time-oriented data described in [2] proposes the following categorization criteria: *time*, *data* and *representation*.

#### Time

Time is a specific data dimension: it could be divided into hierarchical items (days are split into hours, those into minutes, minutes into seconds etc.), it has a cyclic structure (each next day repeats the previous one precisely, so does every week and every season), it also includes social cycles (weekends or holidays, sometimes irregular). Time is common across many application domains. According to its unique features, time-dependent data needs to be treated differently than other kinds of data and require appropriate visual and analytical methods to explore and analyze it [3].

The notion of time has always deserved much attention of researchers. Different classifications and models of time are covered in literature. One of the taxonomies describing different types of time is represented in [2]. The author describes two criteria for classifying time:

- discrete time points vs. time intervals
  This addresses the temporal primitives that the time axis consists of. A time point can be considered an instant in time, or, in contrast to that, it can be specified by two time points or by a time point plus a duration [2].
- linear time vs. cyclic time vs. branching time
  This classification describes the difference in the possible structure of the time axis. Linear time corresponds to our natural perception of time, i.e., time proceeds from the past to the future. A cyclic time axis is composed of a finite set of recurring temporal primitives (e.g., the seasons of the year). On a cyclic time axis, any temporal primitive is proceeded and succeeded at the same time by any other temporal primitive (e.g., winter comes before summer, but winter also succeeds summer). Branching time axes are modeled as graphs, where the vertices with more than one outgoing edge indicate a split of the time axis into alternative scenarios [2].

Discrete time points allow to notice small changes in the data within one day, while time intervals are more suitable for representing aggregated values. For the current project, we are interested in seeing changes, but another concern is not to overload the visualizations with data. That is why we choose data intervals, make them quite small (one minute) and take the median of all values within this interval. As for the way of time representation, linear time is a better choice for the visualization method for the current project, than cyclic or branching time. Cyclic time could be an interesting option if taken apart from the calendar, otherwise it will not fit nicely into the square of one day in the calendar. Thus, time intervals in combination with linear or cyclic time representation will give the proper visualization images.

#### Data

The choice of visualization techniques depends a lot on the data that is necessary to represent. In [2] the following categorization is proposed:

- abstract vs. spatial
  Information visualization, graph visualization, or software visualization are more concerned with abstract data, whereas spatial data are addressed by scientific visualization (flow visualization, volume visualization) or geographic information systems [2].
- univariate vs. multivariate
  The number of variables corresponding to each temporal primitive (a time point or time interval) could be one (univariate data) or more (multivariate data). In the latter case, an additional visualization goal — the detection of correlations — is introduced [2].

- data vs. data abstractions
  For large datasets it is sometimes more reasonable and convenient to derive data abstractions (e.g., to calculate aggregated values).

The desirable number of the dimensions to be seen in our visualization is at least three: life events, time axis and stress level data; in future, the number of dimensions that are represented in the visualization will probably grow, e.g., temperature scale, activities recognized from the accelerometer data, can be also included into the visualization. However, a balance between represented information and visualization readability should be kept: too much information shown in the visualization makes it hard to understand. The amount of data to visualize is going to be huge: even in case a sensor is taking measurements with the frequency of 10 Hz only during working hours, a weekly dataset can reach $1,440,000$ records. It is impossible to build a readable visualization representing all this data, therefore, data abstraction (aggregation) should be done. The extent of data aggregation depends on the aim of a particular visualization: if we want to see all the changes in the stress level in dynamics, the aggregation period should be minimal (e.g., one minute), and the other way around, if the dynamical changes are not important, as opposed to long-term patterns and trends, then the aggregation period can be longer (e.g., one hour, one day).

**Representation**

Two fundamental subcriteria described in [2] are:

- static vs. dynamic
  Static representations visualize time-oriented data in still images. In contrast to that, dynamic representations change automatically over time by means of slide shows or animations.
- 2D vs. 3D
  The number of dimensions to use for visualization of a distinct dataset depends on its size, complexity and kind of data.

Another categorization discussed in [21] is based on form, color, motion and spatial position. Since two last were already described above, we add one more subcriteria to the proposed classification:

- form vs. color
  A form pattern can be represented in a number of ways: line orientation, line form, line width, line collinearity, size, curvature, spatial grouping, blur, added marks, numerosity [21]. A color pattern shows the differences in the values by using hue (different colors) or intensity (different nuances of the same color) [21]. The choice of colors for the visualization is a very important decision. Appropriate use of color for data display allows interrelationships and patterns within data to be easily observed. The careless use of color will obscure these patterns [5]. The approach implemented in the online tool for selecting specific map color schemes [5] describes three kinds of color schemes, presented in figure 2.5.
  - **Qualitative schemes** do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative schemes are best suited to representing nominal or categorical data [5].
  - **Sequential schemes** are suited to ordered data that progress from low to high. Lightness steps dominate the look of these schemes, with light colors for low data values to dark colors for high data values [5].
  - **Diverging schemes** put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues [5].

  In some cases, form and color techniques are combined for better expressiveness.

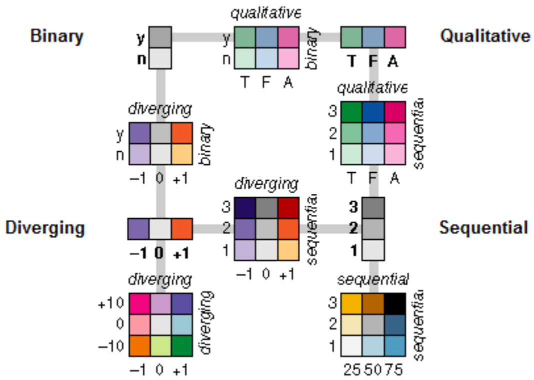



Figure 2.5: Color Scheme (from [5])

In this project, we concentrate on building static visualizations, which represent the data that is already obtained from the sensor. Dynamic visualizations might be possible in the future as drawing pictures in real time based on the data being received from the sensor and transmitted via bluetooth or wi-fi. The number of dimensions in the visualization can be different, but if shown in the calendar, 2D visualization will fit better the square of one day in the calendar. Both form and color patterns look useful and applicable for our visualization, with a couple of constraints. Color scheme looks bright and vivid on the one hand, but on the other its perception depends on the person. Due to non-uniform color reception (not speaking of color-blind persons), this pattern is not always applicable. Besides, color schemes sometimes do not have a clear scale; sometimes it is hard to differentiate between shades of one color. Moreover, the same color can be perceived differently, depending on its surrounding. Given this, it can be hard to define the exact colors corresponding to the extrema. Form pattern sometimes implies traditional graphs, pie/line/bar/dot charts, plots. That eliminates misunderstanding of the scale, but arises another problem: they are not always understandable for people without technical background. However, form pattern helps to notice global trends, which is not possible when using colors.

### 2.2.2 Examples

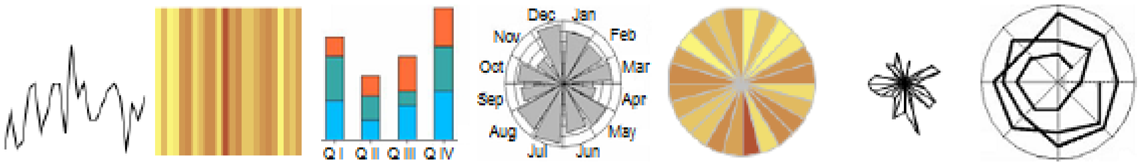
Now we illustrate the classification proposed above with some examples of visualization techniques, combining different values of the discussed criteria.

#### Basic Techniques

Some conventional ways to visualize simple time-dependent datasets are represented in figure 2.6. These graphs are usually static, 2-dimensional in most cases, could use both linear (2.6 (a), (b), (c)) and cyclic (2.6 (d), (e), (f), (g)) time axis representation, implement time intervals (2.6 (b), (c), (d), (e)) or separate time points (2.6 (a), (f), (g)) time pattern and depict only a few variables. These charts are nowadays among the most popular time series visualization techniques, they are widely used in science and business and actively discussed in the literature (see e.g. [14], [8]).

Some of these techniques can be applied to the case of the current project. As it was mentioned above, the ones with the cyclic time representation (Sector and Circle Graphs, Star and Clock Glyphs) are not very convenient for placing into the space inside of the calendar, but can be considered as possible options. The bar chart represents the time axis as intervals, and thus operates with aggregated values. The Line and Stripe Glyphs have an appropriate time concept and are likely patterns for being used in the visualizations.

The charts mentioned above are very common and widely used, but they are not always suitable for more complicated datasets, including multivariate data and longer periods of time, like the




a) Line Glyph (from [8]) b) Stripe Glyph (from [8]) c) Stacked Bar Chart (from [14]) d) Sector Graph (from [14]) e) Clock Glyph (from [8]) f) Star Glyph (from [8]) g) Circle Graph (from [14])

Figure 2.6: Some Conventional Ways of Data Visualization

dataset that has to be visualized in this project. However, in most cases, these visualization ways become the basis for more complicated techniques.

### ThemeRiver

A popular way to present time-dependent data is to use special visual metaphors [14]. One of such metaphors introduced in [9] is ThemeRiver, a technique designed for representing multivariate data.

This technique was initially developed for thematic visualizations over time within a large collection of documents. ThemeRiver uses variations in width to represent variations in strength or degree of representation. The frequencies of the occurrence of special words in documents are represented on a continuous linear time axis. The horizontal flow of the river represents the flow of time. Colored currents that run horizontally within the river represent themes. Each vertical section of the river corresponds to an ordered time slice. The vertical proximity of the river currents makes it easy for users to judge the relative width of currents and thus the relative strength of the themes. Similarly, symmetry of the “river” around the horizontal axis makes it easy for users to understand the general patterns and trends [9].

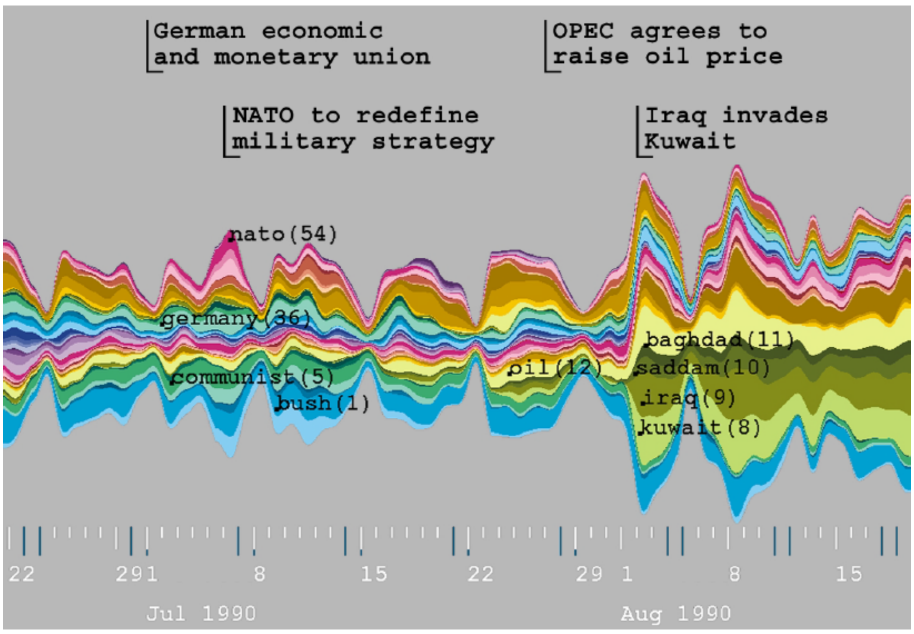



Figure 2.7: ThemeRiver focusing on the theme changes on Iraqi invasion of Kuwait (from [9])

Figure 2.7 shows an example of this technique representing the data of Associated Press from July-August 1990. A wide current in the river indicates heavy use of a topic, while changes in color distribution correlate to changes in themes.

ThemeRiver technique is an interesting idea for the visualization of our dataset, since it is able to show changes in data dynamically. It represents linear time axis, which provides its easy integration in the calendar. However, ThemeRiver is dedicated for showing general patterns and trends during long periods of time, so it uses aggregated values for time intervals. In the limits of this project, ThemeRiver cannot be used for daily stress visualization integrated in the calendar, but can be used for visualizing the changes in the stress levels during all measurement period (if each signal sample is assigned to one of stress level groups (e.g., low, medium, high), then the size of each group can be counted and the changes of the size of each stress level group can be visualized).

### Spiral

Another example of visual metaphor suitable for visualizing periodic multivariate data is presented in [6]. The main feature of spiral concept is using cyclic time axis representation. It supports many display options: 2D and 3D versions, showing the values with colors, bar charts, circle size and other variations. A special application is developed for easy customizing all described settings. One of the possible visualization alternatives is shown in figure 2.8.

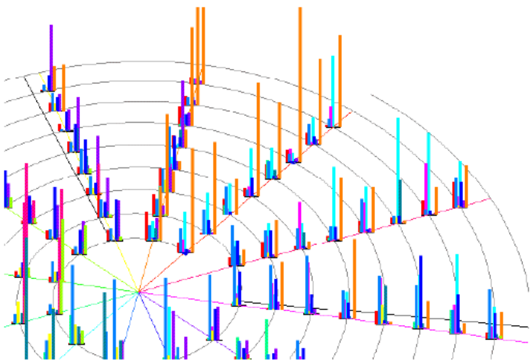


Figure 2.8: An example of Spiral Visualization (from [6])

The stress level visualization described in [16] also uses spiral concept. This system explores a circular perspective of time allowing users to view their activities separated from standard units. In this interface the big centre sphere shows data in real-time. This data is then translated into the history that grows outwards in a spiral. With this interface, it is easy to compare different states over time since they are shown in parallel cycles: data from the previous seconds, minutes, hours, or days are placed in the spiral (with a maximum of three cycles in the spiral). By doing so, users can start comparing and finding patterns in different parts of their data. They can squeeze the data to condense it (compress data), or stretch it to see more detailed information [16].

For our project, the spiral concept is nice for showing the changes in measurements in dynamics (for example, with the color pattern). But on the other hand, cyclic time axis representation limits the possibilities of placing the visualization into the calendar.

### Calendar View

One more obvious and intuitive visual metaphor is calendar view. A calendar visualization presented in [10] concerns e-mail habits of a person (checking, sending, responding, etc). This technique enriches a standard Outlook calendar view by month with colored bar charts (see figure 2.9).

This visualization method provides a possibility to notice dynamically changing stress level data. A bar chart shown for every day allows to follow the changes in the visualized data within one day. An unusual feature of this visualization technique is horizontal orientation of the linear time axis. The advantages of the commonly placed vertical time axis are more convenient arrangement of stress level data together with personal events and simpler process of comparison the visualizations of the same days of the week.

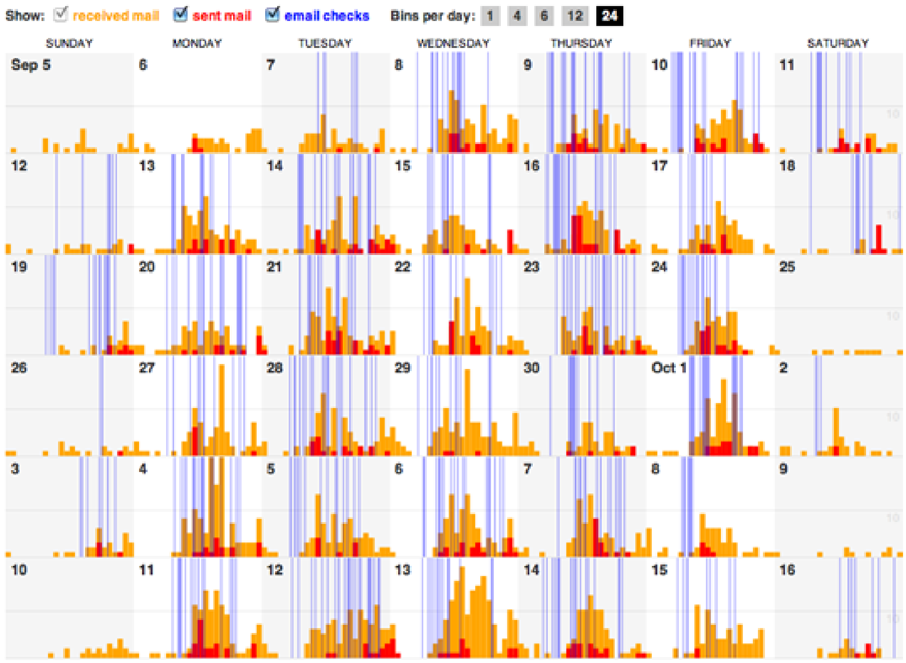



Figure 2.9: Example of Calendar Enrichment with a Bar Chart (from [10])

### Heatmap

In the field of calendar visualization, which implies enriching the space taken by one time entity (day, week, month) in the calendar with some additional visual data (pictures, colors, shapes, charts, etc), there is some variety of techniques, which should be used with care. Since the calendar with personal events represents additional visualization dimension, patterns used along with should be simple and easily understandable. One of the most popular way to enrich the calendar coloring a rectangle representing one day. This approach is used in the heatmaps — graphical representations of data where the individual values contained in a matrix are represented as colors [7], initially introduced for the aims of cartography. Based on this idea, the Calendar Heat Map [18] was designed as a reusable piece of code in programming language R. The result of application of this is presented in figure 3.9. Similar technique is used in the car crashes investigation [24], which was later extended in [15]: each day in the calendar is colored according to the aggregated value of the variable (in this case, number of car crashes).

The heatmap is an interesting idea of visualization of a whole month or a year. This technique uses aggregated values per day, so exactly one color is assigned for one day, which makes impossible to show any changes in data in dynamics. Instead of that, the heatmaps help to notice patterns in stress level data and global trends within one month or one year. In this project, we concentrate on the showing the changes in the stress level within one day, but in future this technique probably can be implemented for the month and year calendar views.

Summing up, there is a lot of work dedicated to visualization of different kinds of data, including stress level. Various approaches to data visualizing discussed above serve different purposes and emphasize different things in the images created. Most techniques use several visualization dimensions, using different means of representing them (e.g., color, shape, size, orientation, etc.). A specific class is calendar visualization, where calendar itself is a particular dimension enriched

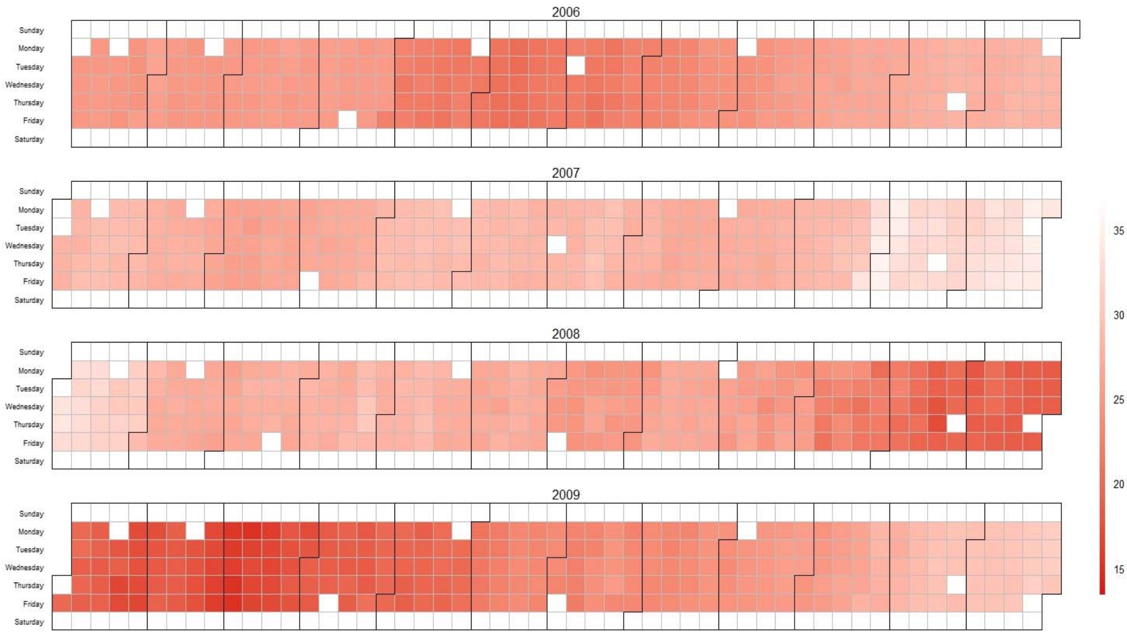



Figure 2.10: Example of Calendar Heatmap (from [4])

by some data. All discussed visualization techniques demonstrate their advantages and drawbacks in accordance with visualization aims defined in chapter 1. Some of the introduced techniques contain interesting ideas that will be used in the development of the new visualizations.

## 2.3 Tools

For various visualization goals there are some tools helping in creating charts and graphs.

One of the most well-known and widely used is MS Excel. It provides a lot of possibilities for creating plots, bar and pie charts, histograms by means of an integrated tool. It also allows heat maps creating. The main advantage of the integrated Excel tool is its clear and intuitive interface, so that the user does not need any knowledge of a specific programming language for using it. However, this reflects in an obvious disadvantage: the tool includes only a limited number of available chart types, serving only main business purposes. ColorBrewer [5] is a specific web tool for selecting color schemes for thematic maps, most usually for choropleth maps, a very useful and easy in use tool based on the color way of data representation.

For adjustable and adaptive visualizations other tools can be more suitable. One of them is SAS/GRAPH, an additional graphic component for the SAS programming language, developed by researchers of SAS Institute [1]. SAS/GRAPH provides users with an efficient, versatile and intelligent means to get a handle on large volumes of multivariate data and to create multidimensional graphs. Usage of this tool requires some programming knowledge. R is another programming language allowing to create visualizations, used for developing Calendar Heat Map [18]. Popular object-oriented languages such as Java and C# also have their own libraries providing a wide range of opportunities for visualization. E.g., JFreeChart library can generate pie charts, bar charts (regular and stacked, with an optional 3D-effect), line charts, scatter plots, time series charts (including moving averages, high-low-open-close charts and candlestick plots), Gantt charts, meter charts (dial, compass and thermometer), symbol charts, wind plots, combination charts and more.

# Chapter 3

# Visualization Ideas

In this chapter we try to adjust various visualization techniques described in chapter 2 to the purpose of the current project. Each section presents one visualization idea and proposes some variation. At the end of the chapter, all ideas are compared to each other in terms of information capacity.

## 3.1 Independent Visualizations

In this section, we discuss visualization techniques that use the information about life events taken from the personal calendars, but themselves are not embedded into a calendar.

### 3.1.1 Pie Chart Visualization

Pie chart is a very popular visualization technique for showing relative sizes of data pieces. In this project, can be an interesting option to look at the structure of a day of a person. For example, for some people several days of week have a similar schedule, some people have meetings at some fixed interval of time (weekly, monthly, etc.). It could be interesting to use pie charts for comparing the days with the same structure to each other in order to find some patterns or trends in the stress level. The color schemes described in figure 2.5 can be used for choosing the colors for the sectors of circles.

Another option of using pie charts could be grouping the data by a type of activity labelled to some period of time. In this case, it would be possible to observe, how stressful are different kinds of activities in general, based on the whole observation period. Figure 3.1 shows the pie charts of the stress level grouped by activities. Diverging color scheme is chosen for visualizing example data.

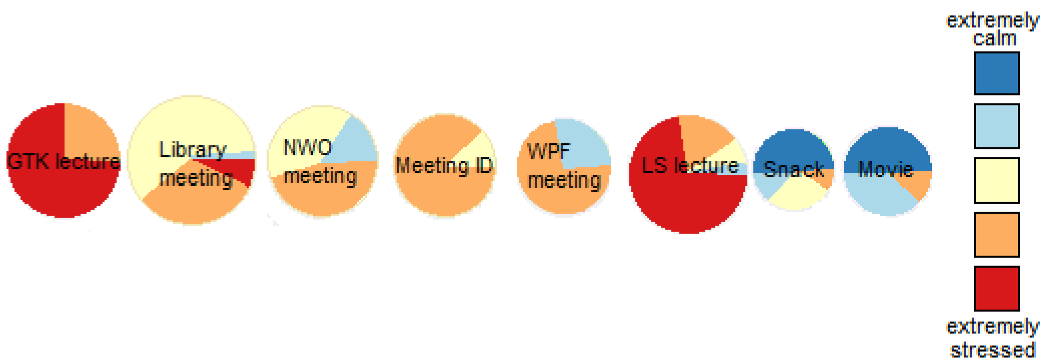



Figure 3.1: Pie chart visualization

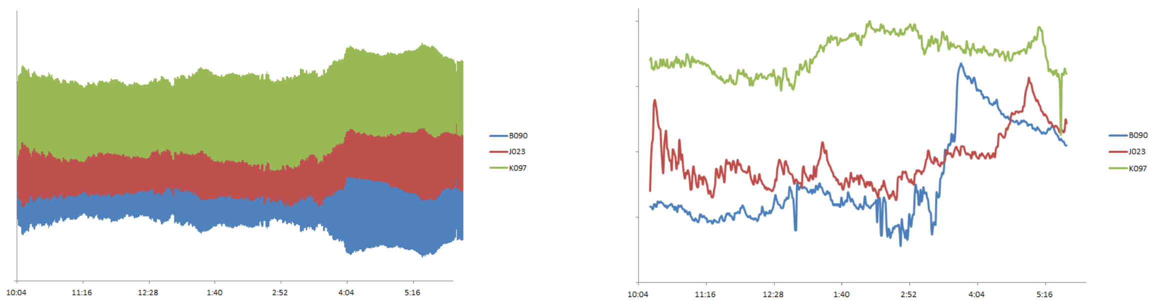
Provided examples of pie chart visualization take the stress level data in isolation from time dimension, therefore, they cannot be put into a calendar. However, this method is informative if there is much information about life events and activities of a person.

### 3.1.2 ThemeRiver Visualization

Theme river is an interesting visualization technique providing rich opportunities for our case. The following possible applications of this technique can be proposed:

- multiple people being together attending the same event (e.g., considering a weekly meeting within the same working group, compare how stress of several people changes with the course of time, if there are any similarities in the way of changing the stress level of different people in the same conditions)
- the multiple instances of the same kind of event of the same person (e.g., considering an everyday event, look if there are some stress patterns for the concrete event type).

Figure 3.2 (a) shows a possible way to represent the first kind of situation: the horizontal axis shows the time flow, and each flow represents the stress level of one person. As an alternative for the ThemeRiver visualization, figure 3.2 (b) shows a commonly used line graph. For building both visualizations, real stress data of three people has been used.




a) using ThemeRiver technique

b) using a line graph for each person

Figure 3.2: Visualization of the stress level of three people attending the same event

Comparing these two visualization methods, we can see that the ThemeRiver visualization can be confusing: a peak in the blue flow spreads to other flows and creates a wrong impression. The same time point in the line graph clearly demonstrates a peak in only one line. ThemeRiver makes it impossible to see individual data in the entire picture, although it excludes flows intersection, which may be hard to understand in the line graph. Summing up, none of two discussed visualization methods is suitable for the aims described above.

### 3.1.3 Spiral Visualization

The stress level data we have to visualize is periodic, since days have similar structure. For expressing the periodicity pattern, the spiral visualization is a nice idea, since it brings the advantage of aligning stress level data for days or weeks. Each time interval can be marked with a circle denoting the stress level by its size or color. Figure 3.3 illustrates the idea for the stress level data aggregated per hour and using size of the circle to show the stress level.

Another possibility is coloring the whole spiral with a continuous color sequence. The problem in this case can be smaller space available for one day in the middle of the spiral than in the outer part. This can make it hard to distinguish one color from another in the middle of the spiral.

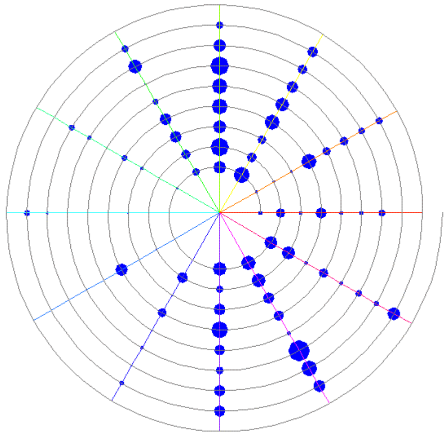


Figure 3.3: Spiral visualization (from [6])

## 3.2 Calendar Visualization

In this section, we speak about different ways of calendar enrichment with stress level data. First, we introduce several techniques for detailed visualization of each day in the calendar in section 3.2.1. Then, in section 3.2.2, we propose some visualization methods for halves of the days, using clock metaphor. Section 3.2.3 provides an idea for visualization stress data for long periods of time, like a month or a year.

### 3.2.1 Day Visualization

#### Color Bars Visualization

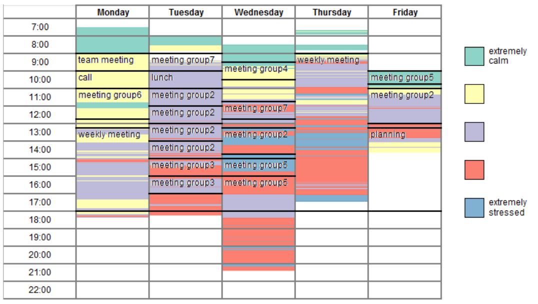
This visualization is built based on the stripe glyph shown in figure 2.6 (b). The time axis is linear, divided into short time intervals of 1 minute; the stress level values are aggregated for each interval, and indicated by the bars of equal width having different colors. Figure 3.4 shows examples this visualization technique with variation on three criteria: color scheme, color scale and color bar width.

**Color scheme**

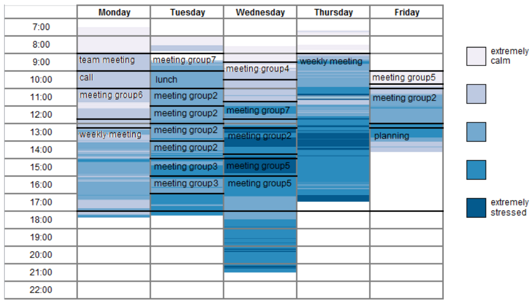
Figure 3.4 (a) shows an example of using **qualitative** scheme for stress visualization (the colors are taken from [5]). Qualitative schemes are the best to represent nominal or categorical data, so using them for stress visualization could make the visualizations hard to interpret, since the hues used are really different, and it is not intuitively clear which color represents a higher stress level and which represents a smaller one. In other words, qualitative color schemes do not provide clear ordering of colors, necessary for fast and unambiguous understanding of the stress level visualizations. As it can be seen from figure 3.4 (a), there is no intuitive understanding of colors ordering, so if there would be no legend, the figure could be interpreted in a completely different way.

As opposed to the qualitative schemes, color ordering is very clear in **sequential** schemes. In the example shown in figure 3.4 (b) all differences in the stress level are shown using different saturation of the same color (in the experiment it was blue). This alternative also has an advantage of colorblind-safety. Having solved a problem of confusing non-ordered colors, we come to the appearance of a new one: the shades are sometimes hard to differentiate, so the picture becomes not quite expressive.

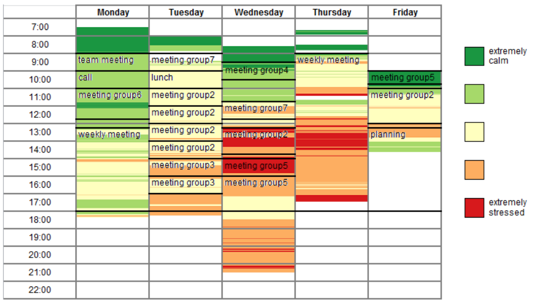
Using a **diverging** color scheme is another possible option for the stress level visualization. In figure 3.4 we show two examples of using diverging color scheme.



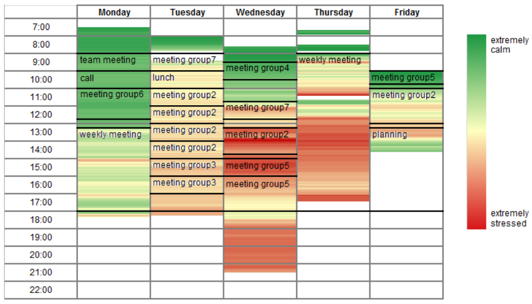

a) qualitative color scheme



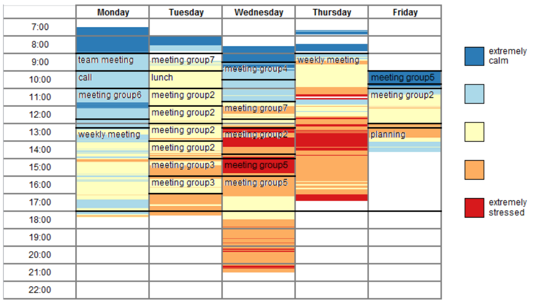

b) sequential color scheme



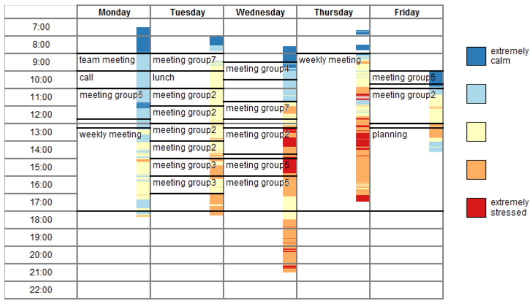

c) diverging color scheme: discrete scale




d) diverging color scheme: continuous scale




e) diverging color scheme: wide color bars




f) diverging color scheme: narrow color bars

Figure 3.4: Color bars visualization

One of the most obvious color choices is **red-yellow-green** pattern shown in figure 3.4 (c). These colors are widely used in all kinds of applications and therefore intuitively understandable by users, for example, the same scheme is used in the traffic lights. Being an advantage on the one hand, intuitive understanding of the color scheme can become a significant drawback on the other hand: green color is definitely perceived as something good, while in the case of stress at work a lot of green means a lot of relaxation, which is not supposed to be normal at work. At the same time, the values meaning common working state lie somewhere close to the medium, which is represented by the hues of the yellow color.

Another solution using the diverging color scheme is shown in figure 3.4 (e). In applies **red-yellow-blue** scheme and thus, excludes the unnecessary association with a traffic light.

Summing up, the diverging color scheme appears to be the most expressive and clear one for this visualization type. Among two alternatives of diverging color scheme proposed in this chapter, the **red-yellow-blue** variant shown in figure 3.4 (e) is the most expressive and not confusing.

#### Color scale

The difference in color scales is evidently shown in figures 3.4 (c) and (d). Both figures apply "red-yellow-green" alternative of the diverging color scheme and use the same width of the color bar. Figure 3.4 (d) uses a **continuous color scale** with 267 colors and a smooth transition from one color to another: the more green a period of time is, the lower values of the stress level it illustrates, and, on the contrary, more red color means higher values of stress level. In other words, in this picture the actual values are just converted to the colors. Figure 3.4 (c), applies **discrete color scale** including only 5 distinct hues without using gradual transition from one to another, thus, one color corresponds to a range of values.

Among these two alternatives, the one using a discrete color scale looks more expressive and clear. A finite predefined number of stress levels allows to notice a distinct difference between them, which is not easy to do in case of too many hues that are hard to distinguish from one another.

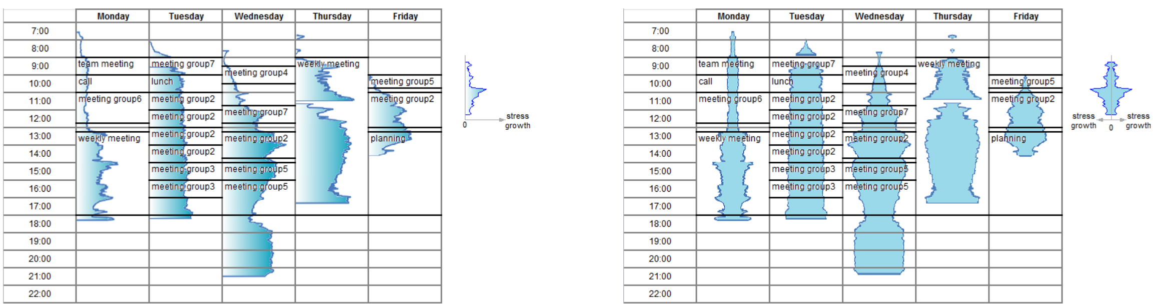
#### Color bar width

The alternatives shown in the figures 3.4 (a)-(e) present a version where the width of the color bars is equal to the width of the day, and in figure 3.4 (f) the colored bars become more narrow to clear the area where the event names are written. Thus, in the alternatives 3.4 (a)-(e) the accent is put onto the visualization itself, which makes the calendar events a minor thing, and the version in figure 3.4 (f) has an advantage of non-intersection of event names and visualization, so the events are better readable.

Both variants have their positive and negative sides, but in this project we want to put the visualization but not calendar events into the center of attention, therefore, the one with wide color bars is chosen.

### Shape Visualization

As a basis of this visualization, the line glyph shown in figure 2.6 (b) is taken. Shape visualization is built for the case of linear time axis, and is represented in the form of an ordinary line graph depicting the stress level data and integrated into the personal calendar. Figure 3.5 shows two possible variations of this technique: 3.5 (a) represents a graph colored from the inside with different color intensity growing as the value grows; figure 3.5 (b) shows the graph filled with single color and reflected on the vertical axis, so it appears to be symmetrical.



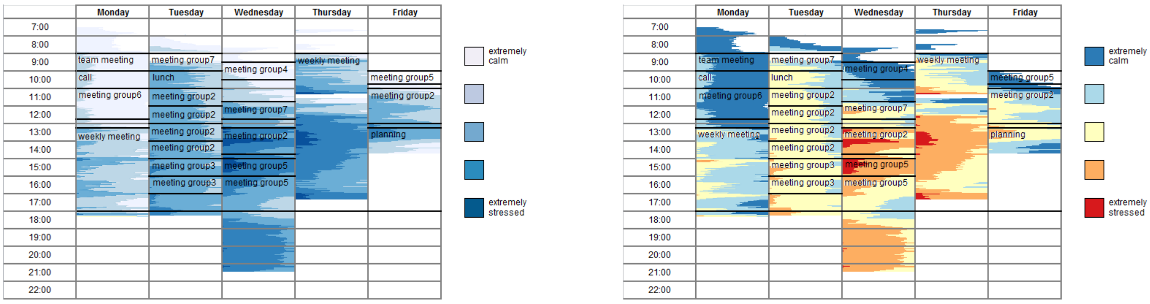

a) gradient colored shape b) symmetrical shape

Figure 3.5: Shape visualization

On the one hand, the peaks, critical values and small fluctuations of the stress level are better seen on the picture (a), but on the other hand, symmetrical shape (especially about a vertical axis) is easier perceived by people [23] and can be clearer for users without a technical background. Between two presented alternatives of shape visualization, the preference is given to the symmetrical one since it is easier to understand [23].

### Colored Shape Visualization

This visualization is based on the technique described in [19]. The idea behind the method is using color and shape at the same time: different stress levels are denoted by different colors (larger values are overplotted in successive color), and gradations within one level are shown by changes in the width of the shape. In figure 3.6 two examples of this technique implementation are shown: (a) uses sequential color scheme, and (b) implements a diverging one.




a) using a sequential color scheme

b) using a diverging color scheme

Figure 3.6: Colored shape visualization based on the technique from [19]

Combining color and shape in one picture makes the image more expressive and informative. Moreover, visualizing stress level data by this method makes the picture look more realistic than the one done by the color bars method, e.g., a bit of red color means a whole red stripe in the color bars technique, but the colored shape technique allowed to show the actual amount of red by using shape. Compared to the shape visualization, colored shape reflects more clearly small fluctuations and big peaks of the stress level due to existence of the color scale.

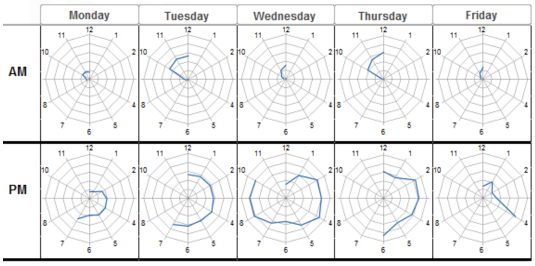
## 3.2.2 Half Day Visualization

### Star Visualization

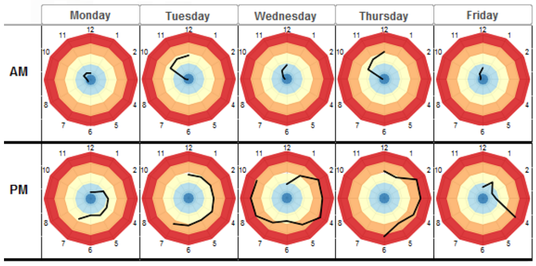
The basis for this visualization is a star glyph (figure 2.6 (f)). The scenarios to be shown with the star visualization can be the same as for pie chart visualization: comparing the stress level of the days with similar schedule, or grouping stress by activity. The star can have s everal axes depending on the number of stress level categories, and the percentage of each stress level category can be shown on the respective axis.

An important advantage provided by the star visualization is its similarity to the clock and presence of a circular time axis. To take this advantage, the star visualization can be used like a clock: a stress level value aggregated per hour (or even a smaller period of time, if does not obstruct the visualization readability) can be shown on the visualization. One day can include two visualizations: one for AM and one for PM. Figure 3.7 shows examples of such visualizations. Figures 3.7 (a) and (b) show the stress level values aggregated per hour and 3.7 (c)-(f) apply aggregation per minute. Alternative (a) only applies the shape pattern, while (b)-(f) also include colors to denote stress levels. Examples (d)-(f) show the visualizations cut by the shape and colored inside. Alternative (d) demonstrates that too low values or too little stress level data for a cetrain day result in a very small picture which is hard to understand. The encountered problem led to creation of the alternatives (e) and (f), showing the color scale for the whole day and only for the periods that have stress data to visualize.

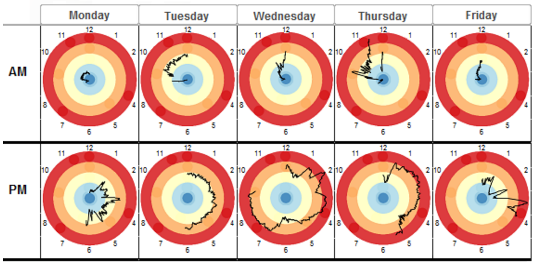
A short opinion poll of ten people proved that figures 3.7 (b) and (c) are the easiest to perceive, but people with technical background also point alternative (f) as giving most clear and detail information.



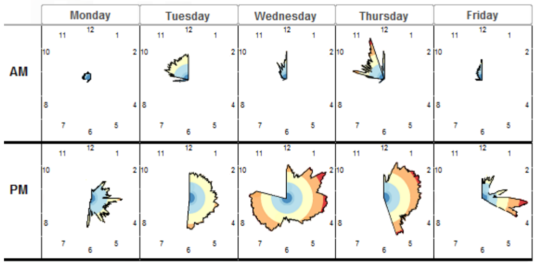

a) aggregation per hour



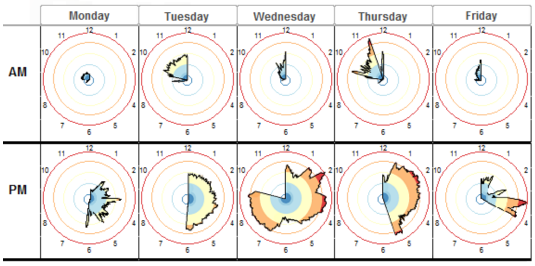

b) aggregation per hour, colored



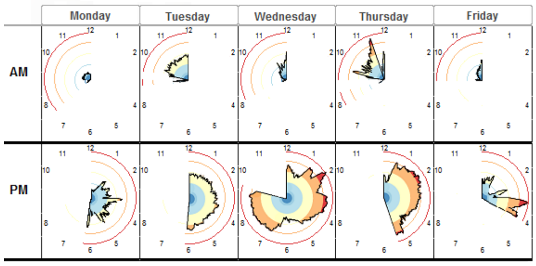

c) aggregation per minute, colored




d) cut by the shape

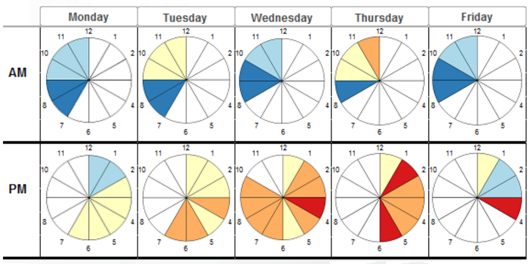



e) color scale for the whole day

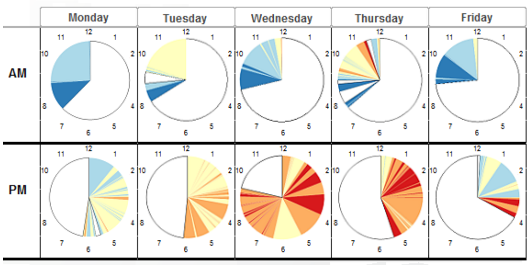



f) color scale for the visualized periods

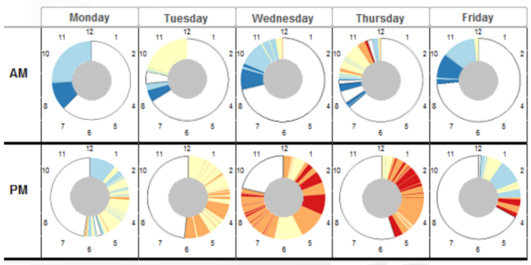
Figure 3.7: Star visualization with the stress value aggregated per minute




a) stress values aggregated per hour




b) stress values aggregated per minute




c) stress values aggregated per minute

Figure 3.8: Clock visualization

#### Clock Visualization

This visualization type, as well as the previous one, has a big advantage of the representation similar to a clock. The difference is that the values should be represented by colors, following the circular time axis. The values can be aggregated for each hour or a smaller period of time, if it does not obstruct the visualization readability. One day can include two visualizations: one for AM and one for PM. Figure 3.8 shows an example of such visualization. Figures 3.8 (a) and (b) show the stress level values aggregated per hour and per minute, respectively. Figure 3.8 (c) is a variation of the alternative (b) without the central part of each circle, where colors are not clear becase of little space.

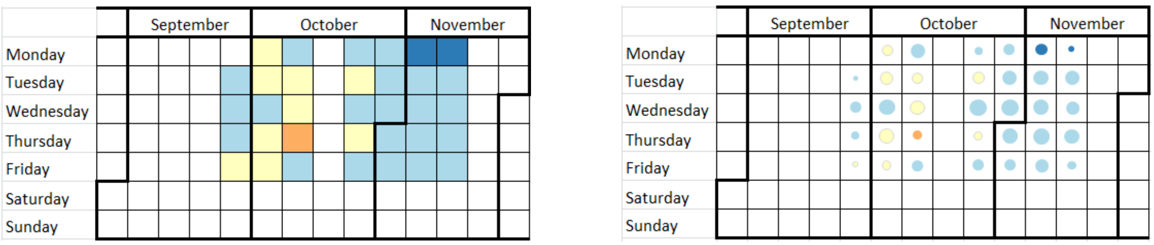
Users without a technical background choose the alternative (a) as the most understandable and clear. Considering all visualization alternatives for half day visualization showed that more participants give their preference to the clock technique than to the star one.

### 3.2.3 Month/Year Visualization

#### Heatmap Visualization

Heatmap can be used to show a long period of time in a calendar (e.g., a month or a year) and coloring each day with a single color. The choice of the color for the day can depend on the mean, median value of the day, the color mode or some other factors. Figure 3.9 (a) shows an example of heatmap technique, visualizing 7 weeks of data.

As a rule, the amount of data available for each day is not the same. In order to make the visualization reflect the real amount of data available for a certain day, it makes sense to color not the whole squares representing days, but to draw a colored circle inside, the size of which will represent the relative amount of data available for the visualization on that day. It can look like shown in figures 3.9 (b).




a) not showing the amount of data available for visualization

b) showing the amount of data available for visualization

Figure 3.9: Heatmap visualization

Figure 3.9 (b) gives more information than figure 3.9 (a): apart from stress level value of the day, it shows the amount of data obtained from the sensor. However, in case of too little stress level data a circle can be that small, that it is not even possible to distinguish its color. This problem can be solved by representing the days having too little stress level data with a circle of a predefined minimal size.

## 3.3 Comparison

Table 3.1 summarizes all visualization ideas described in this chapter and compares them by a number of aspects.

Table 3.1: Visualization ideas comparison

| Aspects | Visualization ideas | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | pie chart | color | | | shape | | color shape | | star | | | | clock | | heatmap | | ThemeRiver | |
| | | qual | div | seq | half | symm | div | seq | flat | hour | min | star | hour | min | days | circles | lines | river |
| **time** | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| perception | | discrete | | | continuous | | continuous | | discrete | | continuous | | discr. | cont. | discrete | | continuous | |
| time axis | | linear | | | linear | | linear | | cyclic | | | | cyclic | | linear | | linear | |
| aggregation | | min | | | min | | min | | hour | | min | | hour | min | day | | min | |
| **color** | ✓ | ✓ | ✓ | ✓ | | | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| color scheme | div | qual | div | seq | grad. | | div | seq | | div | div | div | div | div | div | div | | |
| **shape** | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | ✓ | ✓ |
| shape type | | | | | closed | closed, symmetr. | closed | closed | open | open | open | closed | | | | | open | closed |
| **size** | ✓ | | | | | | | | | | | | | | | ✓ | | ✓ |
| **events** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | |
| **person** | | | | | | | | | | | | | | | | | ✓ | ✓ |
| dimensions in total | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 2 | 3 | 3 | 3 | 2 | 2 | 2 | 3 | 3 | 4 |

# Chapter 4

# User Studies

This section describes two experiments involving potential users of the visualizations proposed in the chapter 3. Section 4.1 describes a small user study aimed at validating three visualization techniques for a day. The second experiment discussed in section 4.2 includes the elaboration on the visualization for a month by the analysis of users' perception of the three validated visualization techniques.

## 4.1 Validation

The main purpose of this experiment was validation of the day visualization techniques. Three visualizations were compared to each other and analyzed in terms of understandability by users and kind of information that could be learnt from each one.

### 4.1.1 Experiment Design

The method used in the experiment was benchmark task method, which basically means asking participants to answer some questions based on quick visualizations analysis. In our case, participants were asked to look at the pictures of the visualized data and perform some simple tasks. Our experimental factors were visualization type and the questions asked.

The choice of the visualization techniques was due to the presence of the calendar and the most detailed stress level representation in comparison to other methods. Among the visualizations described in chapter 3 three were chosen:

- color visualization;
- shape visualization;
- colored shape visualization.

The tasks to perform on visualizations were of several types:

- peak/pit detection, e.g., finding the most/least stressful events;
- trend detection, e.g., pointing out the moments of the steepest stress level growth, and choosing the days demonstrating the biggest/smallest diversity of stress levels.

At the end, participants were asked to choose the preferable visualization method.

#### Data

The data used for building visualizations was obtained from a wearable physiological sensor in a long-term study (6 weeks). For the experiment, we used 3 weeks of data for building each visualization type for each week. All images also contained fictitious information about daily activities of a person.

**Participants**

There were 24 people (12 men and 12 women) participating in the experiment. The participants were mainly recruited from the student population. All participants had normal or corrected-to-normal vision and did not report color blindness. Their age ranged from 22 to 57 years (median age 25). Each participant had at least finished high school, 14 held a Bachelor‘s, 10 a Master‘s degree. The academic background of the participants was mainly in engineering area.

**Procedure**

The questionnaire was designed as a Word form, so each participant was able to perform the experiment on his own. To obtain unbiased answers for all visualization techniques, we had to exclude the situation when a person answers the same question on a different visualization of the same period. In order to do that, we divided the participants into three groups. Each group was exposed to three different visualization types applied to the data of three different weeks.

### 4.1.2 Results Analysis

The analysis of results showed that people understand the notion of the stressful period in a different way. When users were asked to point out the most stressful period of the day, most of them chose the actual moment when the stress level reaches its peak, but several users showed the whole period of the stress level growth till the highest point.

The opinions about the least stressful periods also differ. Most participants consider short periods of very low stress level, usually in the beginning of the day, as the least stressful moments. A few participants have another point of view: they think that stable medium-low stress level that does not demonstrate noticeable changes during some period of time is a sign of one of the least stressful moments.

Figure 4.1 reflects the number of different answers given to the questions about the most (a) and the least (b) stressful periods. The answers are collected for one of three datasets in relation to the visualization type. In both questions, users were allowed to indicate one or two periods. For some periods represented in the figure, a range of answers denoting approximate time intervals was aggregated and considered as the same one.

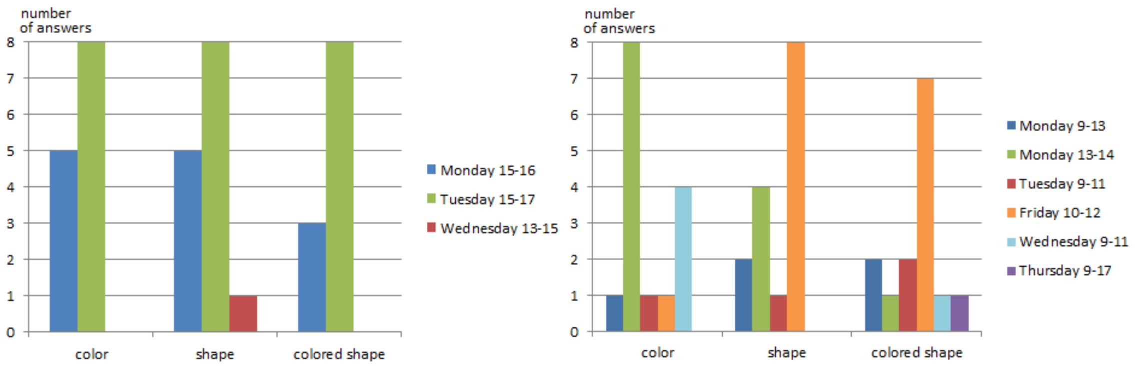
As it can be seen from figure 4.1 (a), users agree on the most stressful time period regardless of the visualization type. The second most popular answer is also the same for all three visualization types. Figure 4.1 (b) demonstrates a bigger variety of users' choices compared to figure 4.1 (a). Moreover, the most popular answer changes depending on the visualization.

Figure 4.2 shows the color variation within the time periods indicated by users as the most (a) and the least (b) stressful. For each interval, the average, maximal, minimal colors and the standard deviation of the color were computed. The colored bars cover the interval $(\bar{x} - \sigma; \bar{x} + \sigma)$, where $\bar{x}$ is the average color within the stated interval, and $\sigma$ is the standard deviation. The purple line with markers denotes the minimal, average and maximum colors in the interval.

Figure 4.2 (a) clearly shows that the most popular answer for all three visualization types (blue bar) demonstrates the higher average and minimum colors, which allows to consider it as the correct one. Since two answers were allowed to give, the second most popular answer (green bar) is also counted as correct because of the most proximity to the right one. Both periods in the figure demonstrate the same maximum color.

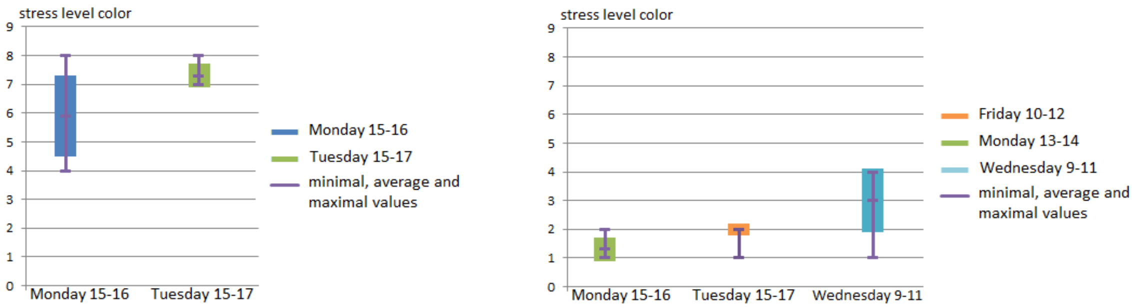
From figure 4.2 (b) it can be seen that the answers that appeared to be the most popular in figure 4.1 (b) indeed demonstrate the lowest (green bar) and the second to lowest (yellow bar) stress levels. The third answer (blue bar) se periods demonstrate close values of the stress level, so we cannot talk about the mistake of the users. The results obtained can be explained by the pictures that the users were exposed to. In color visualization the color abstraction is used, corresponding one color to several stress values. This makes the borders between stress levels very clear, and since the number of colors is not big, people can easily distinct between them and choose the one denoting the lowest stress level. In figure 4.3 we can quickly define two spots of dark blue color, denoting the lowest stress level, on Monday and Wednesday, which are the answers given by the users. In the shape visualization, we do not use any abstraction and visualize real data aggregated per minute. That requires more thinking from users, and when comparing close values of the stress level to each other, an imprecise choice can be made. In figure 4.4 we can see that it is indeed difficult to compare the periods of low stress on Monday and Friday. In case of colored shape visualization, the color abstraction is used together with the shape pattern, so in this case, as well as in the shape visualization, the users had to compare the form of two shapes of the same color. In figure 4.5 we can see that the period on Friday shows bigger amount of dark blue color than Monday, which means bigger stress values, but it could be intuitively understood by users as lower stress, or simply attract more attention than Monday as a colored area.

Figures 4.6–4.7 show the number of different answers given to the questions about the most (a) and the least (b) stressful periods, for two other datasets. These figures demonstrate the tendencies similar to the described ones.




a) the most stressful period b) the least stressful period

Figure 4.1: Dependence of the answers on the visualization type




a) the most stressful period b) the least stressful period

Figure 4.2: The color variation within the time periods indicated in the answers

Two tasks were dedicated to finding the days with the biggest and smallest difference in stress level. Figure 4.8 reflects the number of different answers given for all three visualization techniques for one of three datasets. For these tasks, two answers were allowed as well.

If we explore the data used for building visualizations, we can easily discover the actual range of the stress level for each day. This information is represented in figure 4.9: (a) reflects the range of colors representing stress levels, and (b) shows the range of values in percentage of the biggest possible range. From figure 4.9 (a) it can be seen that two days mentioned by the users as having the biggest difference in stress levels are chosen correctly. Moreover, in color visualization, these days have an equal range of the stress levels, and therefore both answers are counted as correct.

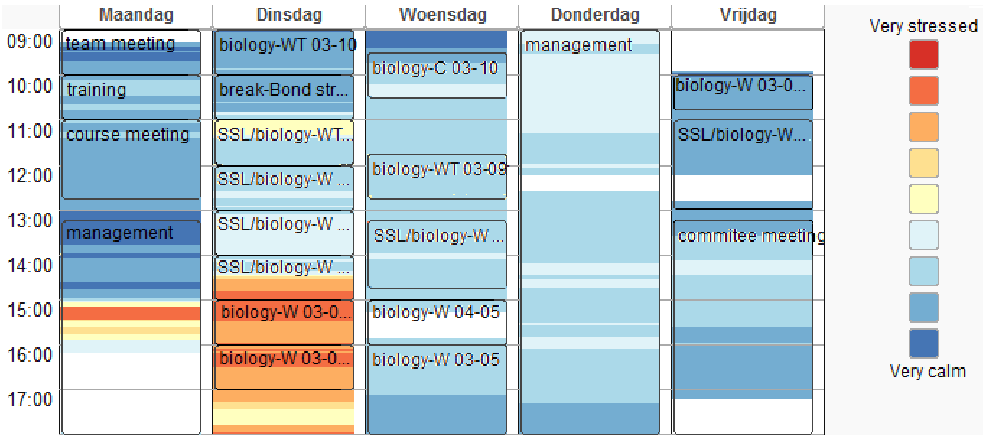



Figure 4.3: An example of the picture used in the evaluation protocol. Color visualization

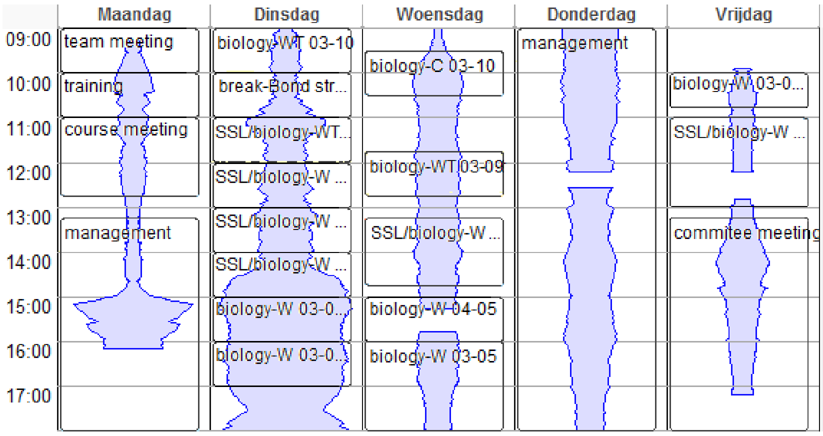



Figure 4.4: An example of the picture used in the evaluation protocol. Shape visualization

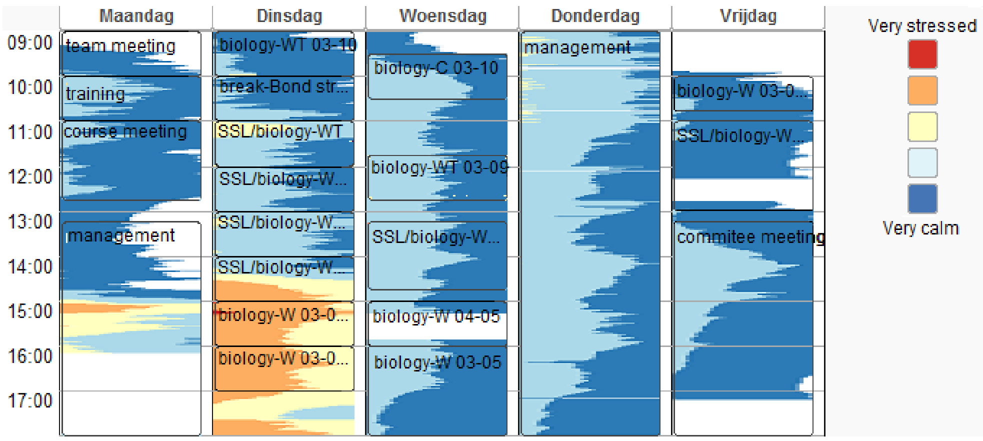
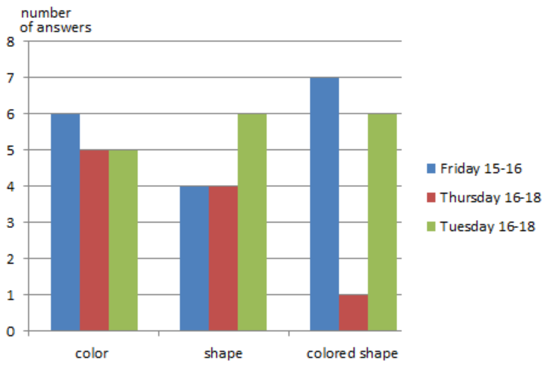
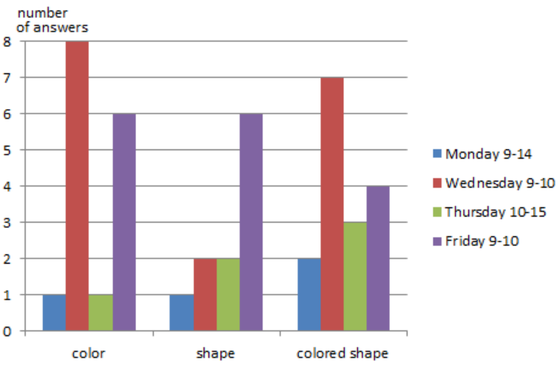



Figure 4.5: An example of the picture used in the evaluation protocol. Colored shape visualization

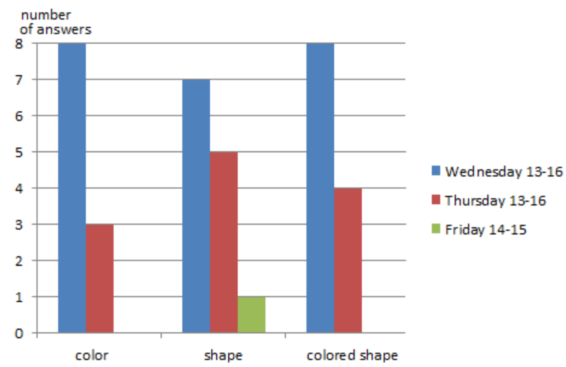



a) the most stressful period

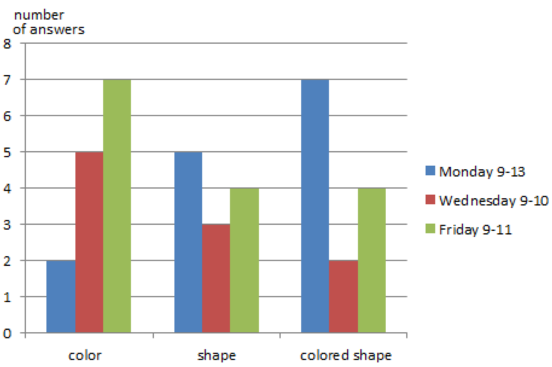



b) the least stressful period

Figure 4.6: Dependence of the answers on the visualization type




a) the most stressful period




b) the least stressful period

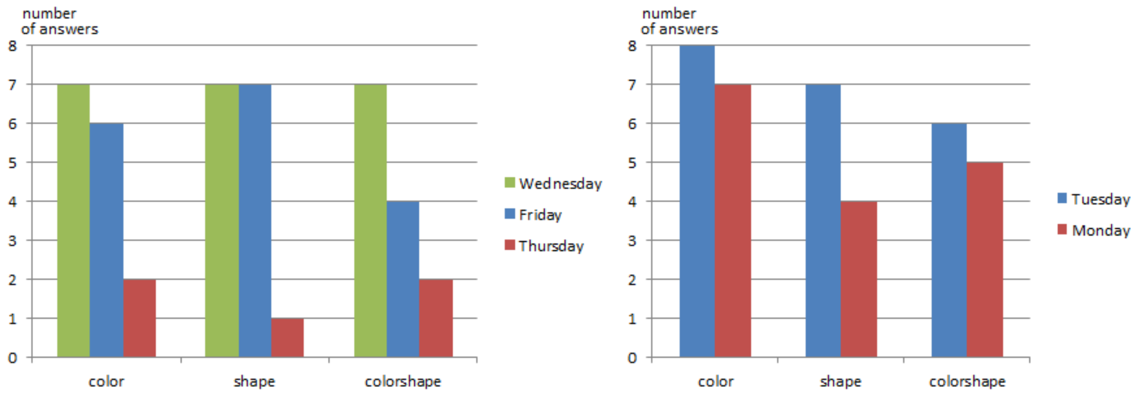
Figure 4.7: Dependence of the answers on the visualization type

For shape and colored shape visualizations, Wednesday shows slightly bigger difference between the highest and lowest points (see figure 4.9 (b)).

In the question about the smallest difference between stress levels, the most popular answer is correct, and the second most popular answer is the closest to the correct answer, which should be considered as correct since two answers are allowed to give. However, from figure 4.9, it is clear that the error of considering Monday as the least stressful day is approximately double. Monday and Thursday in fact do not demonstrate either smallest or biggest difference in stress level. Moreover, they both show equal (for colors) or almost equal (for actual values) difference of the stress level. I believe that the users are confused because in average Monday demonstrates lower stress values than Thursday (see figure 4.9 (a)), making people think that the difference between the lowest and the highest is smaller, though in reality it is not true. Remarkably, people get their answers biased regardless of the visualization type.

The correctness of the answers given by users to these two questions is shown in figure 4.10. The correct answers are represented with blue color, and the blue dotted pattern shows the percentage of the answers that should be considered as correct, but in fact demonstrate an error in comparison with the correct ones.

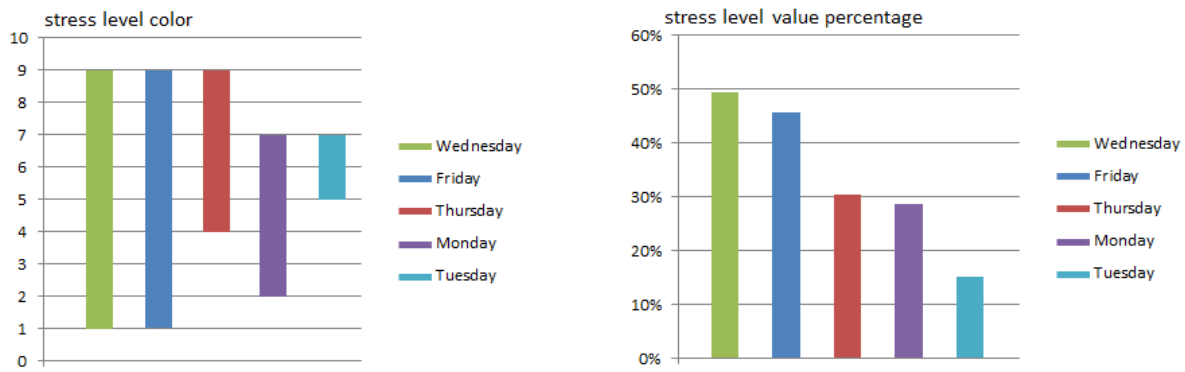
Figure 4.11 reflects the correctness of the answers to these two questions summarized for all three datasets. For both types of questions, color visualization shows the highest percentage of the choices of the "totally correct" answer. From figure 4.11 (a) it can be seen that the best performance in choosing two correct answers is shown by shape visualization type (around 90% correct answers). Colored shape visualization demonstrates twice more wrong answers given by users. Shape visualization shows the worst performance in detecting the day with the smallest




a) the biggest difference between the highest and the lowest stress level

b) the smallest difference between the highest and the lowest stress level

Figure 4.8: Dependence of the answers on the visualization type




a) the range of color representing the stress level

b) the range of values in percentage of the biggest range

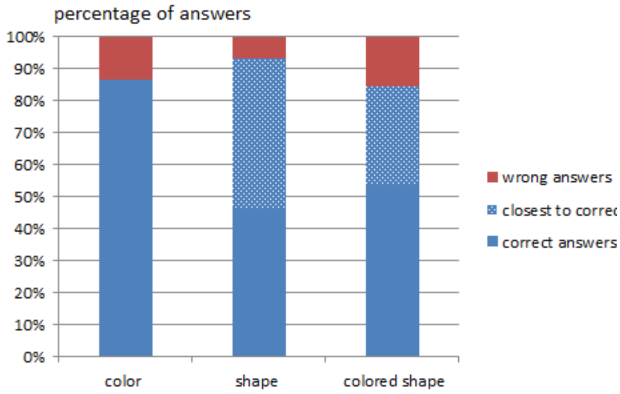
Figure 4.9: The range of the stress level for each day

difference between stress levels (see figure 4.11 (b)) (less than 80%).

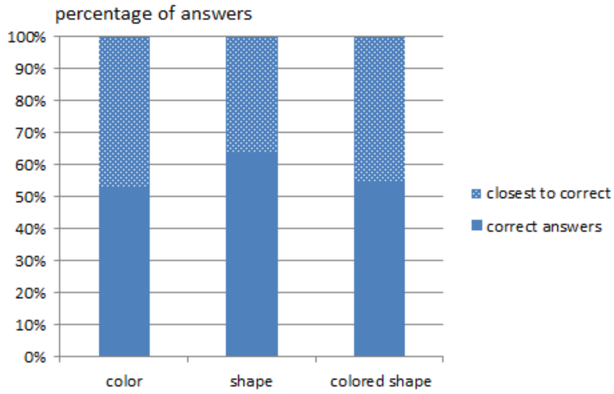
The results of the trend detection task demonstrate correspondence between the visualization types, with a slight better performance for the shape visualization. In this case, color abstraction obstructs clear perception of a trend.

Figure 4.12 reflects the visualization preferences of the participants. According to figure, most users (16) would choose color visualization to be their default one, a quarter of them (6) would prefer to use shape visualization, and only 2 users would like to work mostly with colored shape visualization. Remarkably, out of 8 users who did not choose color visualization as their favorite, 7 have a strong technical background and 7 are male. Colored shape visualization technique turned to be the most disliked one: 9 users reported that they would never use it. Within this group, 5 users would prefer color visualization and 4 would choose shape visualization by default. Out of two users giving their preference to the colored shape visualization, one would never work with color visualization.

Figure 4.13 illustrates the results of hierarchical cluster analysis of the participants based on their preferences of the visualization type. Clusterization was performed using Ward's method with the number of clusters limited up to 3. The desired number of clusters was chosen according to the number of visualizations available to choose as default. The expected result was joining all users having the same preferences about the visualization type in a single cluster.

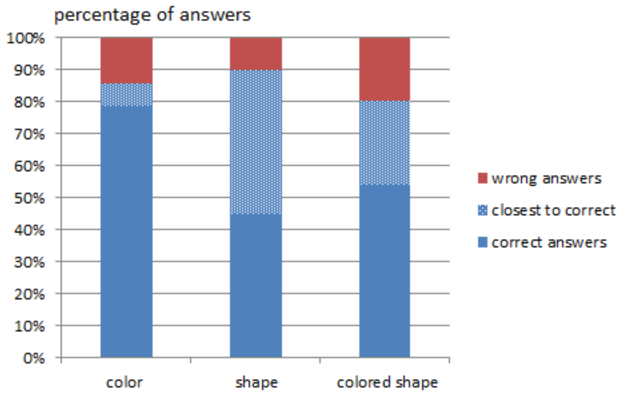



a) the biggest difference between the highest and the lowest stress level

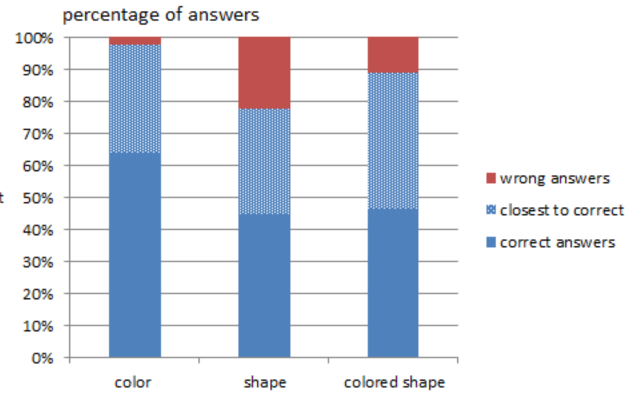



b) the smallest difference between the highest and the lowest stress level

Figure 4.10: Correctness of the answers




a) the biggest difference in the stress level




b) the smallest difference in the stress level

Figure 4.11: Correctness of the answers summarized for three datasets

Each of three clusters joins together the users who have chosen a particular visualization type by default. The first cluster includes 6 participants. All users who of this cluster chose shape visualization type as the default one. None of participants of this cluster ever specified the desired frequency of usage for any visualization as “often”. One third of them would never use color visualization and two thirds said the same about colored shape visualization. The second cluster is the smallest one. It includes only two participants who agreed on having colored shape visualization type as the default one. The third cluster unites the users who prefer to use color visualization. The structure of answers for two other visualization types is similar: approximately the same percentage of people chose “often”, “sometimes” and “never” for shape and colored shape visualizations.

One more conclusion was derived from personal interactions with participants during the experiment. If not explicitly stated, absence of visualization is perceived by some users as extremely calm state, while in reality it means absence of sensor data to visualize. This can be a strong reason to reconsider the denotation of absence of data for visualization.

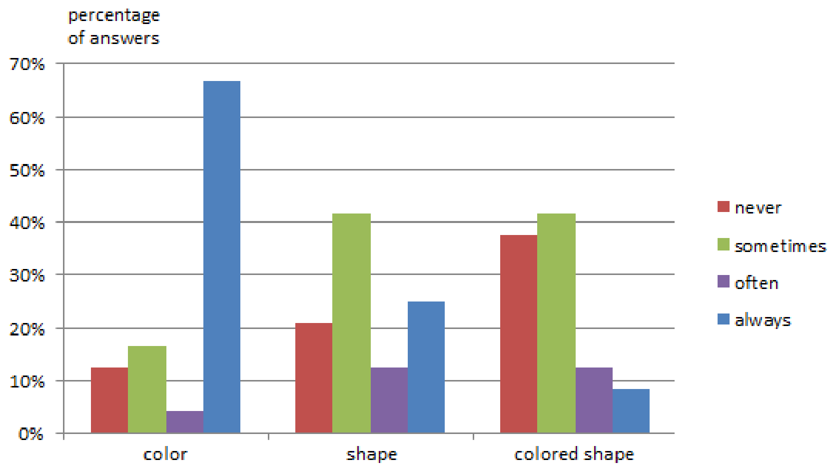



Figure 4.12: Desired Frequency of Visualizations Usage

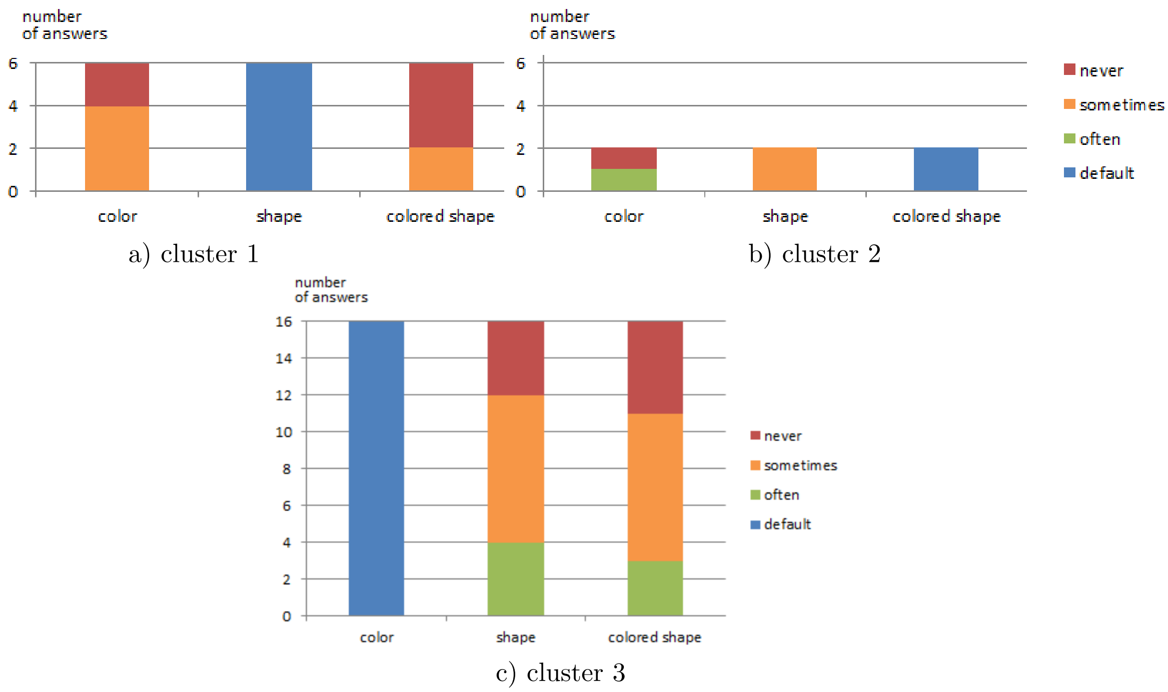



Figure 4.13: Results of the users clusterization

## 4.2 Heatmap Creation

The main purpose of the user study was to get an insight of the way of creating a heatmap visualization for a month, described in chapter 3. To create it, an idea of color definition for an entire day should have been discovered. The experiment performed allowed to learn how people understand the notion of the color of an entire day, what they pay attention to and what they are motivated by when choosing the color characterizing one day.

### 4.2.1 Experiment Design

The method used in the experiment was benchmark task method, which basically means asking participants to answer some questions based on quick visualizations analysis. In our case, participants were asked to look at the pictures of the visualized data and to estimate the stress level for each day using a proposed color scale. Each participant performed the same task 30 times (six weeks of five days each) and for each of the three visualization types.

Our main experimental factor was visualization type. Among the visualizations described in chapter 3 three were chosen:

- color visualization;
- shape visualization;
- colored shape visualization.

The choice of the visualization techniques was due to the presence of the calendar and the most detailed stress level representation in comparison to other methods.

#### Data

The data used for building visualizations was obtained from a wearable physiological sensor in a long-term study (6 weeks). Some images also contained fictitious information about daily activities of a person.

#### Participants

There were 36 people (24 men and 12 women) participating in the experiment. The participants were mainly recruited from the student population. All participants had normal or corrected-to-normal vision and did not report color blindness. Their age ranged from 23 to 55 years (median age 26). Each participant had at least finished high school, 15 held a Bachelor‘s, 20 a Master‘s degree. The academic background of the participants was mainly in engineering area.

#### Procedure

The questionnaire was designed in the form of a web-page, so each participant was able to perform the experiment on his own. The results were collected to an Excel spreadsheet.

### 4.2.2 Results Analysis

#### Preselected Features

The perception of the amount of stress for the whole day based on a visualization of stress measurements during the day may differ from person to person. Different techniques make different things visible, and different people pay attention to different things, e.g., the number of peaks or the amount of some color. One will select the stress level color for the whole day according to a certain metrics that is usually used unconsciously. The question of interest here is if one of the metrics defines the choice of a person; in other words, to understand how the choices of the stress level color for the whole day are related to the features described below:

- the **mean value** of stress during a particular day
  The mean value can be calculated using the following formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{4.1}$$

  where $n$ is a number of minutes, for which the signal exists (sometimes the signal is lost, so there is no visualization for these periods), and $x_i$ is the stress level value for the $i$-th minute.

- the **median value** of stress during a particular day
  The median is the numerical value separating the higher half of the data set from the lower half. This metrics is the most resistant, having a breakdown point of 50%: as long as no more than half the data is contaminated, the median will not give an arbitrarily large result. The principle of using this metrics is the same as for the mean value. The median of a set of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one (in case of even number of observations, there is no single middle value; the median is then defined to be the mean of the two middle values).

- the **standard deviation of the mean value** of stress during a particular day
  Another statistical metrics that could influence evaluating the level of stress during some day is the standard deviation. The criterion here is the opposite: having bigger values of standard deviation mean bigger changes in the stress values, thus, the more stressful days. However, small values of this metrics cannot help to drive any conclusions: it could mean not only that there were small values of stress, but also that there was big stress values with small fluctuations around. The metrics is calculated as follows:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{4.2}$$

  where $n$ is a number of minutes, for which the signal exists, $x_i$ is the stress level value for the $i$-th minute, and $\bar{x}$ is calculated using the equation 4.1.

- the **maximum and minimum values** of stress during a day
  This metrics can be used in a sense that the bigger maximum value it shows during the day, the more stressful the day is. However, using the maximum or minimum value to make the conclusions about the value of the whole day can be ambiguous, since these values can be outliers, so these metrics can be useful in combination with other ones.

All the same metrics can be also calculated for the colors used in the visualization using the same formulas:

- the **mean color** of a particular day

- the **median color** of a particular day

- the **standard deviation of the mean color** of a particular day

- the **maximum and minimum color** of a particular day

- the **color mode** of a particular day
  Mode is the value that appears most often in a set of data. This metrics can be a valuable alternative for the mean or median color.

### Dependence of the answers on the gender of participants

Since men and women have different color perception, it was decided to check if the answers given by these two groups of participants differ. That was checked by means of nonparametric Mann-Whitney. Answers for the two visualization methods using colors were checked separately.

Figures 4.14 and 4.15 show the line graphs for the average answer given by men and women for the color and colored shape visualization techniques respectively. The black dots denote the days that demonstrate statistically significant difference in the answers given by people of different gender.

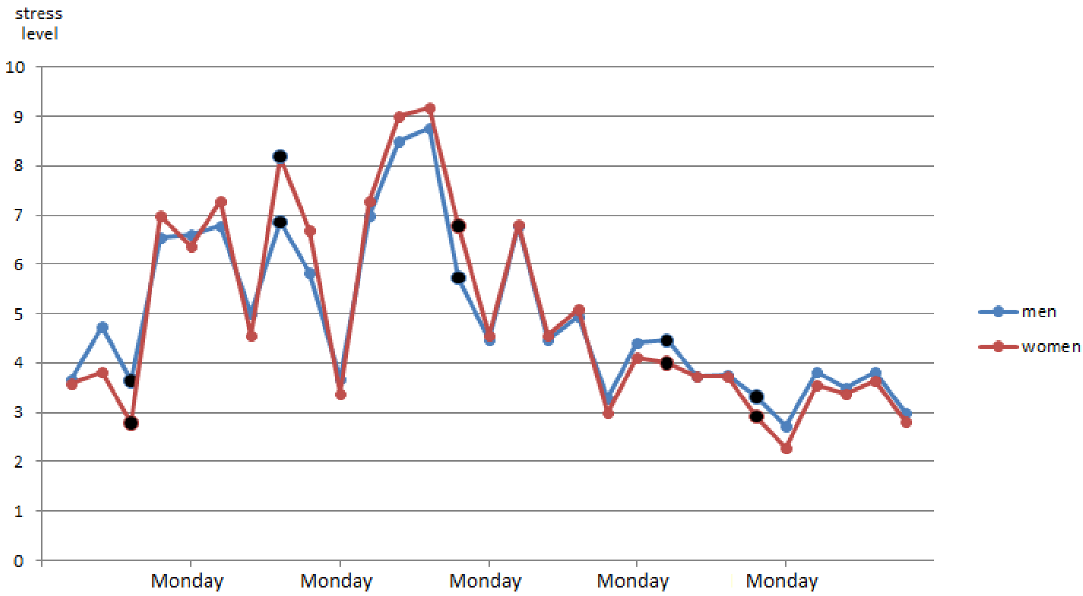



Figure 4.14: Answers given by men and women for the color visualization

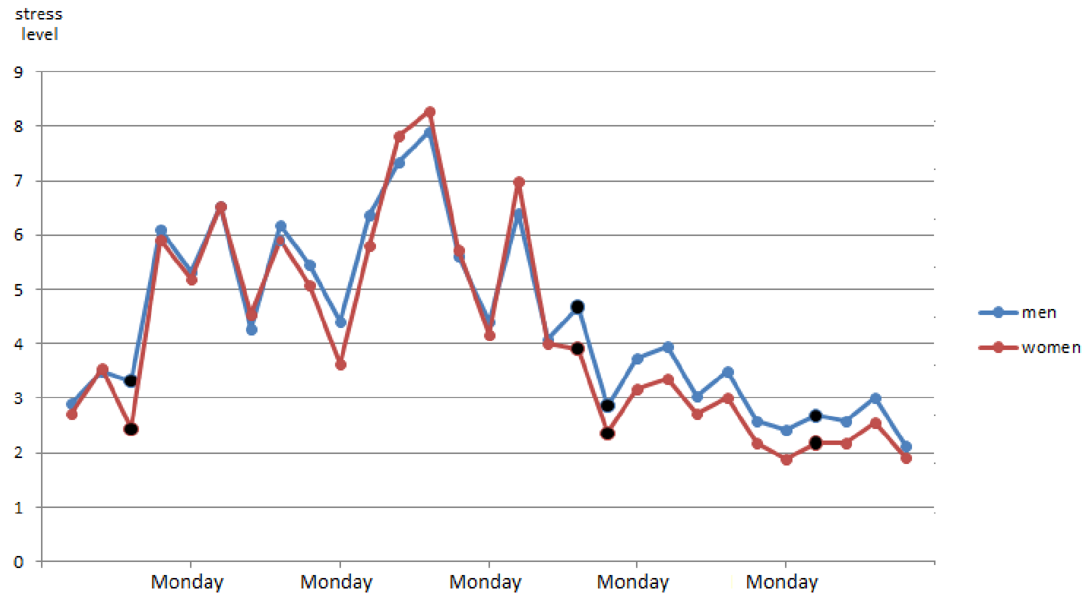



Figure 4.15: Answers given by men and women for the colored shape visualization

For the color visualization, different answers are given for the five out of 29 days. For three of them (Thursday$_1$, Tuesday$_5$ and Friday$_5$), men on the average chose higher color for the stress level of the whole day, and for the other two (Thursday$_2$, Friday$_3$) the average answer given by women was higher. Remarkably, the days estimated by women as more stressful, have more orange and red color and demonstrate bigger standard deviation of the color than the days marked as more stressful by men. That could be a result of a bigger attention of women to the higher values of stress or big standard deviation of the stress value during the day. Besides, the amount of data for two of the days was quite small, which can also be a reason of disagreement between users.

For the colored shape visualization method, men marked four days (Thursday$_1$, Thursday$_4$, Friday$_4$ and Tuesday$_6$) as more stressful than women. These days demonstrate mostly two lowest out of five stress levels. That could be a result of underestimation of lower values of stress level by women.

The analysis of the answers given by men and women for the shape visualization did not discover any statistically significant difference for any of the days. That could mean that the reason of the differences in the answers for two other visualization methods is indeed the inequality of color perception of men and women.

**Dependence of the answers on the visualization type**

In the questionnaire users had to answer each question three times: once for each visualization type. They were not told that in advance, but nevertheless it became clear for many participants. Comparison of the answers given depending on the visualization type represented by a simple line graph is shown in figures 4.16 and 4.17 (the median and the mean answer respectively are taken for each day for each visualization type). In the picture it is clearly seen that the answers to the questions about the same day for different visualizations in most cases vary quite a lot.

In figure 4.18 confidence intervals for the average answer are shown. The whole visualization period of six weeks is split up into two smaller periods, each of three weeks. We can see that for the last three weeks the answers for the shape visualization differ a lot from the answers given to the same question but for two other visualization types. It can be seen that the confidence intervals for the mean for the shape visualization almost never intersect with the confidence intervals for two other visualizations (except for Mondays). Probably it can be explained by the fact that Monday was always the first day to evaluate on all pictures, and participants were investing more effort on correct estimation of Monday, while the rest four days of the week were estimated based on the color given to Monday.

Also, the confidence intervals for the mean answer for shape visualization are bigger than the ones for other visualization types. The table in figure 4.18 shows standard deviations for each day of the visualized period. It is visible from the table that the standard deviations for the last three weeks for the shape visualizations are significantly bigger. This can be explained by the fact that those days demonstrate quite stable stress level with very small fluctuations, and since it does not have any color to be based on, the users had to estimate it intuitively: some of them made it more precise, others — less precise.

Visual inspection of the line graph shown in figures 4.16 and 4.17 and confidence intervals shown in figure 4.18 is not enough to make a conclusion about the presence or absence of influence of the visualization type on the answer given, so the Kruskal-Wallis test was applied.

The result of Kruskal-Wallis test is shown in table A.1. As we can see, p-value$_{\text{Tue}_1} = 0,170 > 0,05$, p-value$_{\text{Fri}_3} = 0,170 > 0,05$ and p-value$_{\text{Mon}_6} = 0,170 > 0,05$, which means that we should accept the main hypothesis about no differences between the answers given for three visualization types for three days. For the rest 26 days, p-value $< 0,05$, so we can reject the main hypothesis and admit that there is a dependence of the answer on the visualization type. However, this only means that there is a difference between at least any of two groups. From table A.2 we can conclude that there is a difference between shape visualization and two other visualization types, but we cannot say anything definitive about the difference between the color and colored shape visualizations. For that reason, we continue our investigation and compare the visualizations pairwise for those 26 days that showed differences between visualization types.

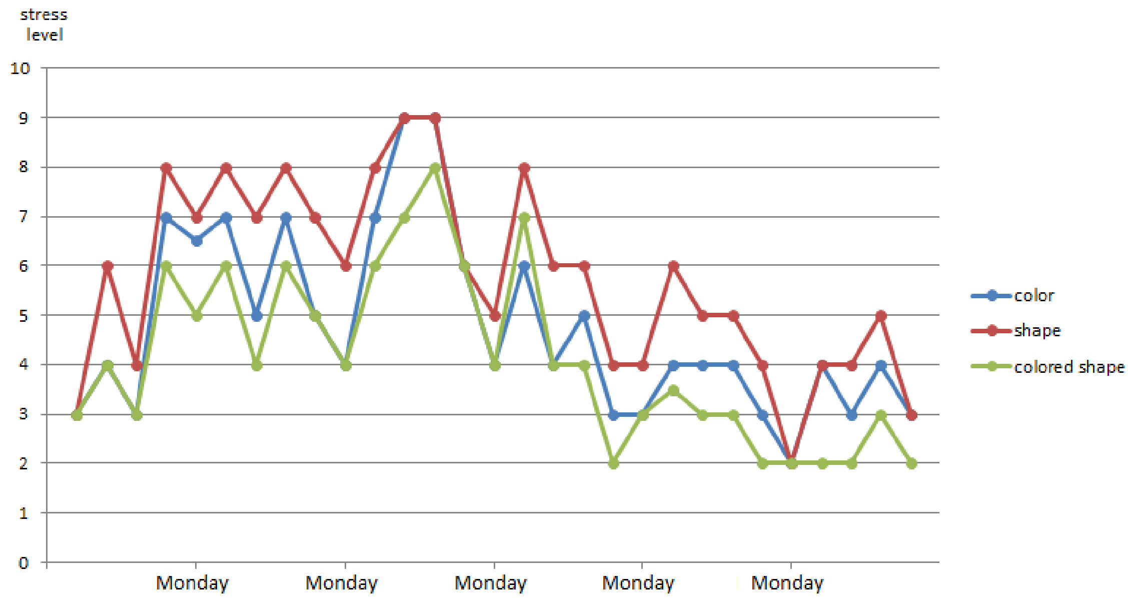



Figure 4.16: Dependence of the answers on the visualization type (median answer)

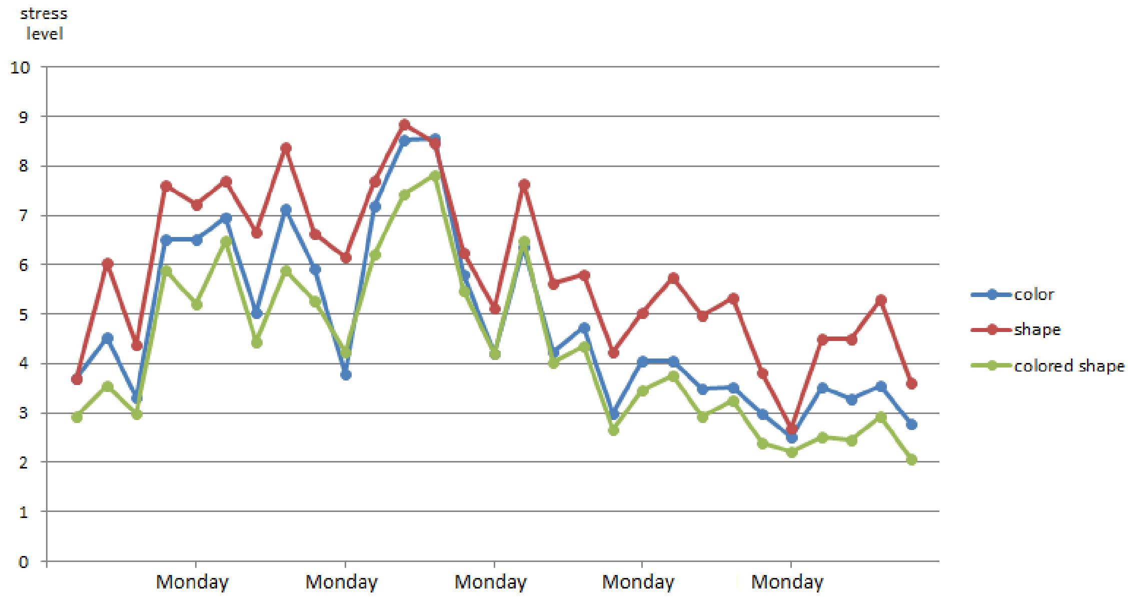
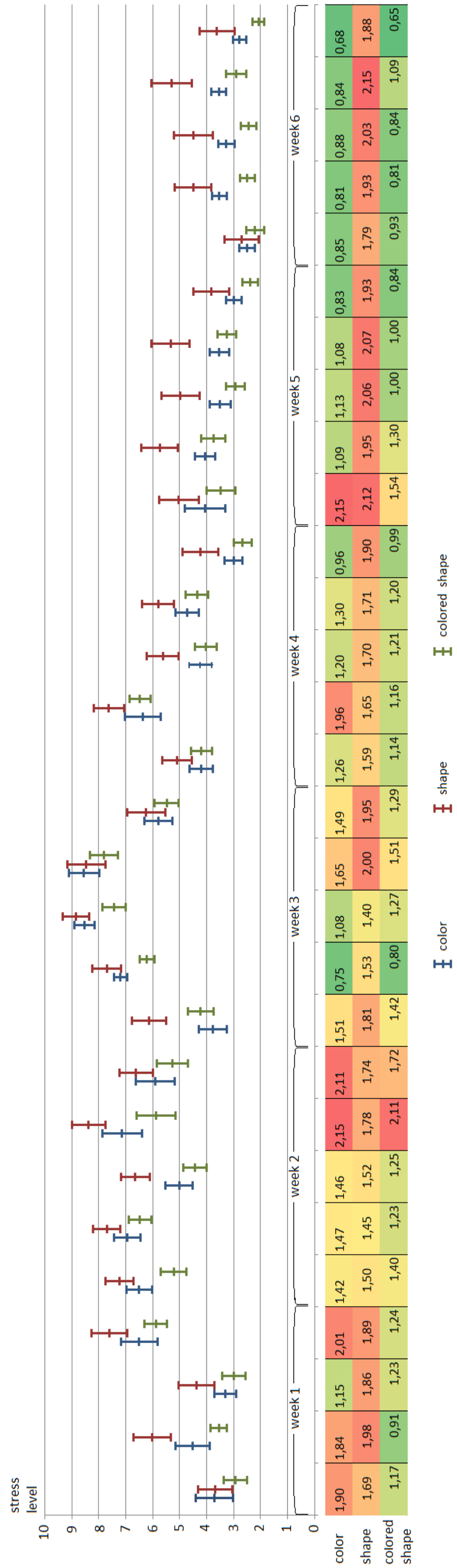



Figure 4.17: Dependence of the answers on the visualization type (average answer)

The results of the pairwise comparisons are shown in table A.1 in the last three columns. The difference between the shape and colored shape visualization types is statistically significant (p-value $< 0,05$ for all 26 days). For most days, there is also a statistically significant difference between color and shape visualization. However, for 6 out of 26 days the difference between the answers for these visualization types are not statistically significant. For around half of the days, the difference between color and colored shape visualization is not statistically significant.




| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| color | 1,90 | 1,84 | 1,15 | 2,01 | 1,42 | 1,47 | 1,46 | 2,15 | 2,11 | 1,51 | 0,75 | 1,08 | 1,65 | 1,49 | 1,26 | 1,96 | 1,20 | 1,30 | 0,96 | 2,15 | 1,09 | 1,13 | 1,08 | 0,83 | 0,85 | 0,81 | 0,88 | 0,84 | 0,68 |
| shape | 1,69 | 1,98 | 1,86 | 1,89 | 1,50 | 1,45 | 1,52 | 1,78 | 1,74 | 1,81 | 1,53 | 1,40 | 2,00 | 1,95 | 1,59 | 1,65 | 1,70 | 1,71 | 1,90 | 2,12 | 1,95 | 2,06 | 2,07 | 1,93 | 1,79 | 1,93 | 2,03 | 2,15 | 1,88 |
| colored shape | 1,17 | 0,91 | 1,23 | 1,24 | 1,40 | 1,23 | 1,25 | 2,11 | 1,72 | 1,42 | 0,80 | 1,27 | 1,51 | 1,29 | 1,14 | 1,16 | 1,21 | 1,20 | 0,99 | 1,54 | 1,30 | 1,00 | 1,00 | 0,84 | 0,93 | 0,81 | 0,84 | 1,09 | 0,65 |

Figure 4.18: Confidence intervals for the average answer and standard deviations of the answers for each day

### Linear regression analysis

In this section we analyze which factors among the ones described in section 4.2.2 influence the choice of participants for one or another visualization type. For each type, different factors may matter, so all three visualizations were analyzed separately. To estimate the accuracy of the predictive models, the method of *cross-validation* was used. To apply this method, the initial dataset was randomly permuted and divided into four parts (folds). One fold was preserved as a testing set, and three others were used for the analysis of the model, performed in three rounds. During each round, the training of the model was done on the set of two folds and validated on one, as shown in table 4.1.

Table 4.1: 3-Fold Cross-Validation Principle

| | $\text{fold}_1$ | $\text{fold}_2$ | $\text{fold}_3$ |
|---|---|---|---|
| $\text{round}_1$ | **validation set** | training set | |
| $\text{round}_2$ | training set | **validation set** | training set |
| $\text{round}_3$ | training set | | **validation set** |

### Color Visualization

For the regression model explaining answers given for the color visualization, the following predictors were chosen:

- the mean color of a day ($\text{AVG}_{color}$),
- the median color of a day ($\text{MED}_{color}$),
- the standard deviation of the mean color of a day ($\text{DEV}_{color}$),
- the maximum color of a day ($\text{MAX}_{color}$),
- the minimum color of a day ($\text{MIN}_{color}$),
- the color mode of a day ($\text{MODE}_{color}$).

In tables A.3 and A.6 the models and validation results for the color visualization respectively for the median and average answers are represented. The last column (error) represents the result of calculation of the metrics proposed for evaluating model quality, in other words, how precise the model predicts the answers of participants. The error is calculates as follows:

$$\text{error} = \frac{\sqrt{\sum_{i=1}^{n} (\text{predictedValue}_i - \text{actualValue}_i)^2}}{n}, \tag{4.3}$$

From tables A.3 and A.6 it can be seen that the main predictor in most cases is $\text{AVG}_{color}$, followed by $\text{DEV}_{color}$. It can be also noticed than the models that include both factors usually demonstrate lower error (on that particular dataset) in comparison to the ones that include only one of them. The correspondence of the actual answers given by users and the values predicted by the models demonstrating the smallest errors is shown in figures 4.19 (for the median answer) and 4.22 (for the average answer).

### Shape Visualization

For the regression model explaining answers given for the shape visualization, the following predictors were chosen:

- the mean value of a day ($\mathrm{AVG}_{value}$),
- the median value of a day ($\mathrm{MED}_{value}$),
- the standard deviation of the mean value of a day ($\mathrm{DEV}_{value}$),
- the maximum value of a day ($\mathrm{MAX}_{value}$),
- the minimum value of a day ($\mathrm{MIN}_{value}$).

In tables A.4 and A.7 the models and validation results for the shape visualization respectively for the median and average answers are represented. The last column (error) represents the result of calculation of the metrics proposed for evaluating model quality, in other words, how precise the model predicts the answers of participants. The error is calculated using the expression 4.3.

From tables A.4 and A.7 it can be seen that the main predictor is always $\mathrm{AVG}_{value}$. Some predictive models also include $\mathrm{DEV}_{value}$, which improves the model performance, and/or $\mathrm{MIN}_{value}$, which makes it slightly worse. It can be also noticed than the models that include both $\mathrm{AVG}_{value}$ and $\mathrm{DEV}_{value}$ usually demonstrate lower error (on that particular dataset) in comparison to the ones that include only $\mathrm{AVG}_{value}$. The correspondence of the actual answers given by users and the values predicted by the models demonstrating the smallest errors is shown in figures 4.20 (for the median answer) and 4.23 (for the average answer).

### Colored Shape Visualization

For the regression model explaining answers given for the colored shape visualization, the following predictors were chosen:

- the mean color of a day ($\mathrm{AVG}_{color}$),
- the median color of a day ($\mathrm{MED}_{color}$),
- the standard deviation of the mean color of a day ($\mathrm{DEV}_{color}$),
- the maximum color of a day ($\mathrm{MAX}_{color}$),
- the minimum color of a day ($\mathrm{MIN}_{color}$),
- the color mode of a day ($\mathrm{MODE}_{color}$),
- the mean value of a day ($\mathrm{AVG}_{value}$),
- the median value of a day ($\mathrm{MED}_{value}$),
- the standard deviation of the mean value of a day ($\mathrm{DEV}_{value}$),
- the maximum value of a day ($\mathrm{MAX}_{value}$),
- the minimum value of a day ($\mathrm{MIN}_{value}$).

In tables A.5 and A.8 the models and validation results for the colored shape visualization respectively for the median and avegage answers are represented. The last column (error) represents the result of calculation of the metrics proposed for evaluating model quality, in other words, how presice the model predicts the answers of participants. The error is calculates using the expression 4.3.

From tables A.5 and A.8 it can be seen that there are two main predictors: $\mathrm{AVG}_{value}$ which is present in all models, and one of the two: $\mathrm{DEV}_{value}$ or $\mathrm{DEV}_{color}$, that always improve models performance. The correspondence of the actual answers given by users and the values predicted by the models demonstrating the smallest errors is shown in figures 4.21 (for the median answer) and 4.24 (for the average answer).

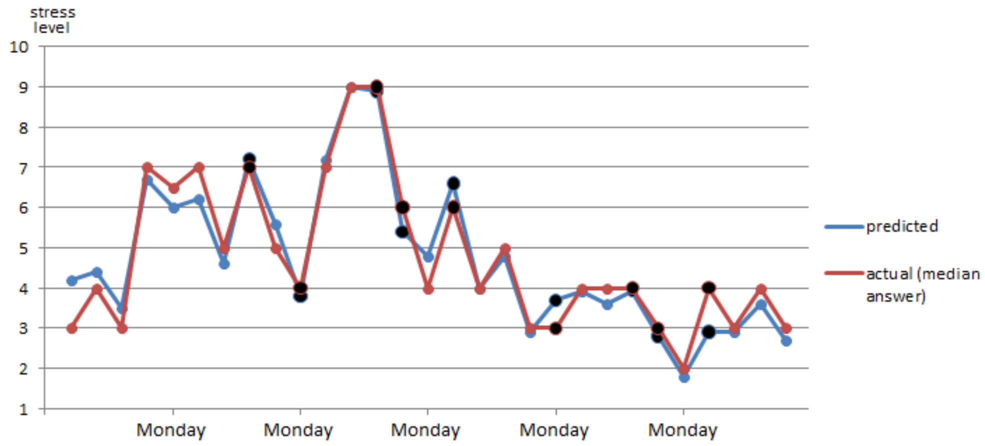



Figure 4.19: Correspondence of the median answer of the users given for color visualization and the values predicted by the model

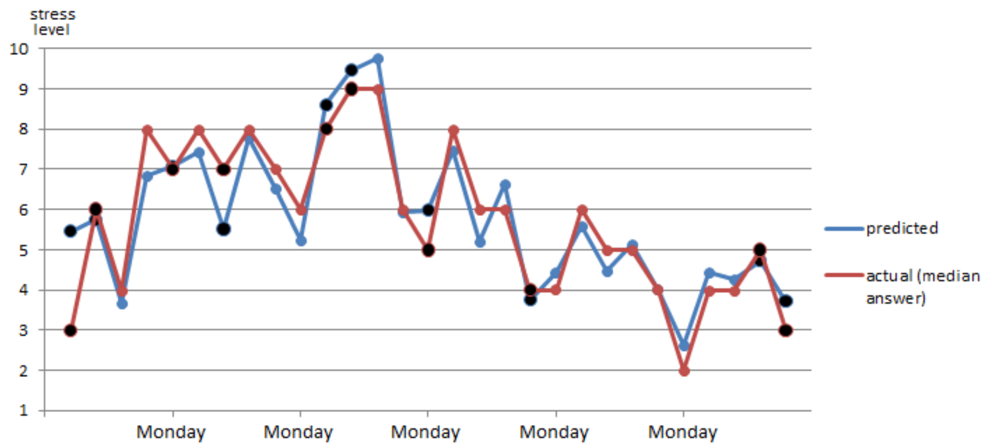



Figure 4.20: Correspondence of the median answer of the users given for shape visualization and the values predicted by the model

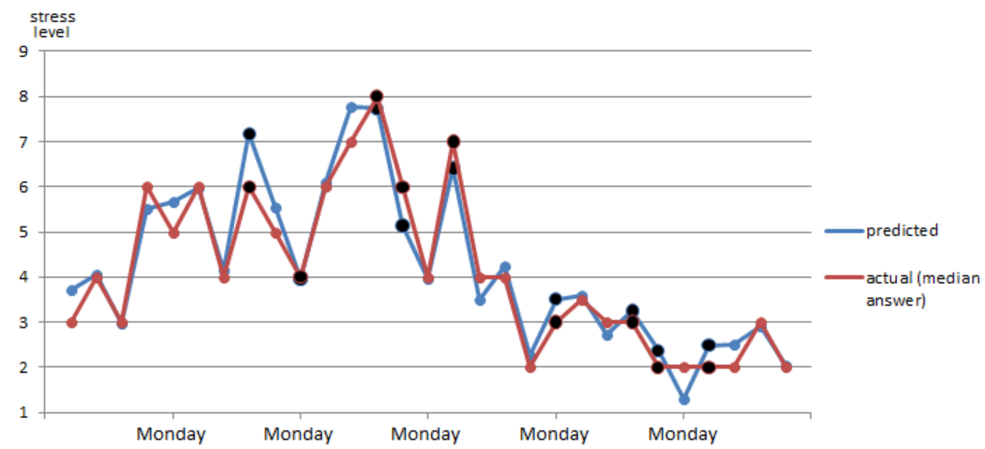



Figure 4.21: Correspondence of the median answer of the users given for colored shape visualization and the values predicted by the model

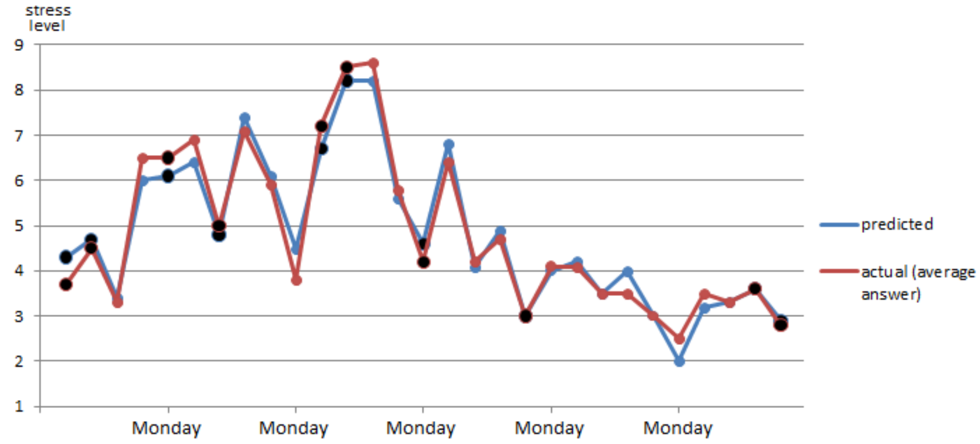



Figure 4.22: Correspondence of the average answer of the users given for color visualization and the values predicted by the model

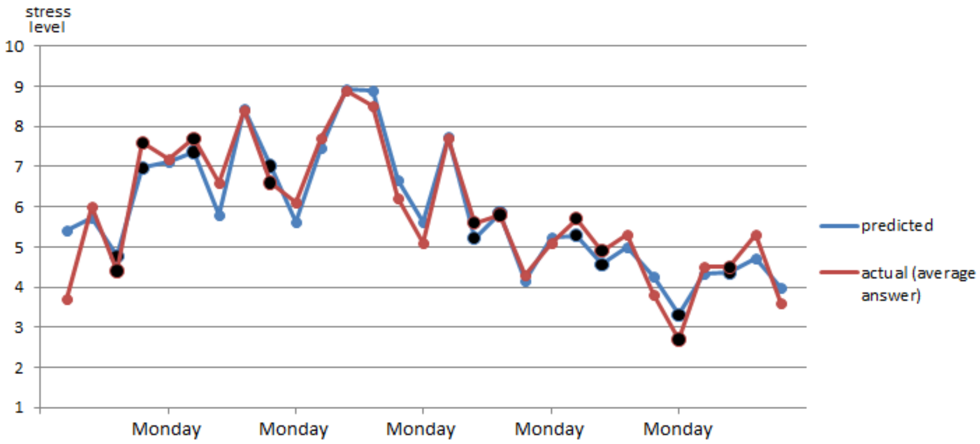



Figure 4.23: Correspondence of the average answer of the users given for shape visualization and the values predicted by the model

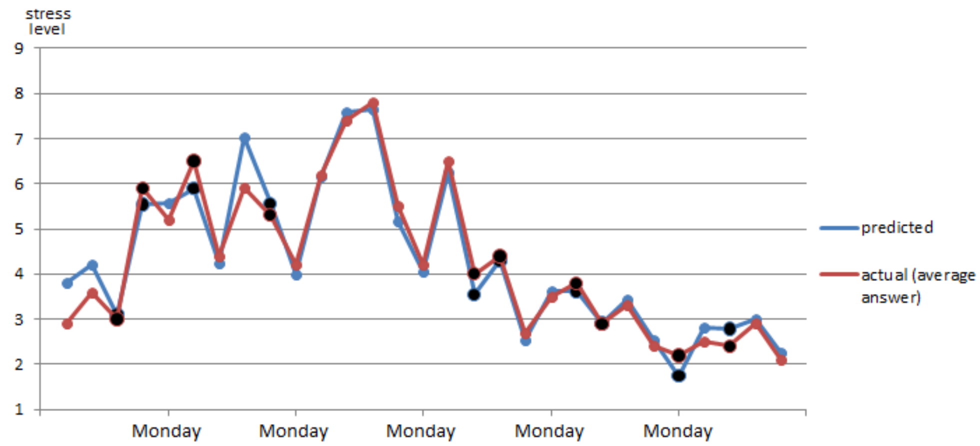



Figure 4.24: Correspondence of the average answer of the users given for colored shape visualization and the values predicted by the model

### Clusterization of the participants

The line graph illustrating the answers of different participants looked very chaotic and showed huge variance between answers, but at the same time some similarities within subgroups of participants. However, just a visual analysis was not enough to draw any valuable conclusions about similarities of some users, so to discover if some participants indeed can be united in groups, hierarchical cluster analysis was hold. The analysis was performed using Ward's method. The number of clusters to get as the result of analysis was limited up to 20.

### Color Visualization

Figures A.1-A.7 demonstrate the results of cluster analysis when the number of clusters was 7. As it can be seen from the pictures, most participants are included into first three clusters (figures A.1-A.3), and four other clusters include one to three participant each. We believe that each cluster unites people using the same motivation while choosing one color for the whole day. Once again regression analysis was used to discover if there is a model predicting the answers for the group (cluster). Only first three clusters were taken into consideration, other four were excluded from analysis as too small. The predictors chosen were the same as for the regression analysis for the color visualization described in section 4.2.2. For each cluster the average answer was calculated and 3-fold cross-validation of the models obtained in the investigation was hold. Table A.9 shows clusterization results for the color visualization. The last column (error) represents the result of calculation of the metrics proposed for evaluating model quality, in other words, how precise the model predicts the answers of participants. The error is calculated using expression 4.3.

From table A.9 it can be seen that each of three clusters indeed has its main predictor different from the predictors of other clusters. That means that users that fall into different clusters pay attention to different features of the visualization when choosing a color of the day. Thus, in the first cluster, which includes 14 out of 34 users and appears to be the biggest, the main predictor is $\text{MED}_{color}$, though in some models $\text{MIN}_{color}$, $\text{DEV}_{color}$ and $\text{AVG}_{color}$ also appear. For the second cluster, including 7 participants, the main predictor is $\text{AVG}_{color}$, followed by $\text{DEV}_{color}$, and in some cases also $\text{MED}_{color}$. In the third cluster, which includes only 5 users, the predictor $\text{DEV}_{color}$ is included in all models, but when it is combined with $\text{AVG}_{color}$ or $\text{MED}_{color}$ the models demonstrate better performance, which is reflected in the smaller values of the error. However, all predictors characterize only the average answer of the cluster.

Figure 4.25 shows the popularity of the predictors in formulas describing actual models for separate people within the three biggest clusters.

As can be seen from the figure, each cluster has one or two main predictors. For the first cluster, the most popular one is $\text{AVG}_{color}$, followed by $\text{MED}_{color}$. The main predictor for the majority of users (57%) is $\text{AVG}_{color}$, and only for 29% of users of this cluster it appears to be $\text{MED}_{color}$. According to the results of the linear regression analysis of the average answer, the main predictor for this cluster is $\text{MED}_{color}$. I believe that the such a difference in the results can be caused by using the average answer of the cluster to find out the predictors, which is imprecise due to the big number of participants in the cluster.

For the second and the third clusters, the results obtained match the conclusions made from the linear regression analysis of the average answer for each cluster. In the second cluster, the most popular predictor is $\text{AVG}_{color}$, which is the main predictor for 71% of users of this cluster. In the third cluster, the $\text{DEV}_{color}$ is the main predictor for 80% of users.

Summarizing all the conclusions made above, even having the same visualization type, different people choose the color of the day in a different way. Based on the answers participants give, they can be united into some groups (clusters) according to the features they consider important when choosing the color of the whole day. However, even with a big number of clusters, the differences within them are quite noticeable. That is why this investigation can be continued only with a bigger number of participants.

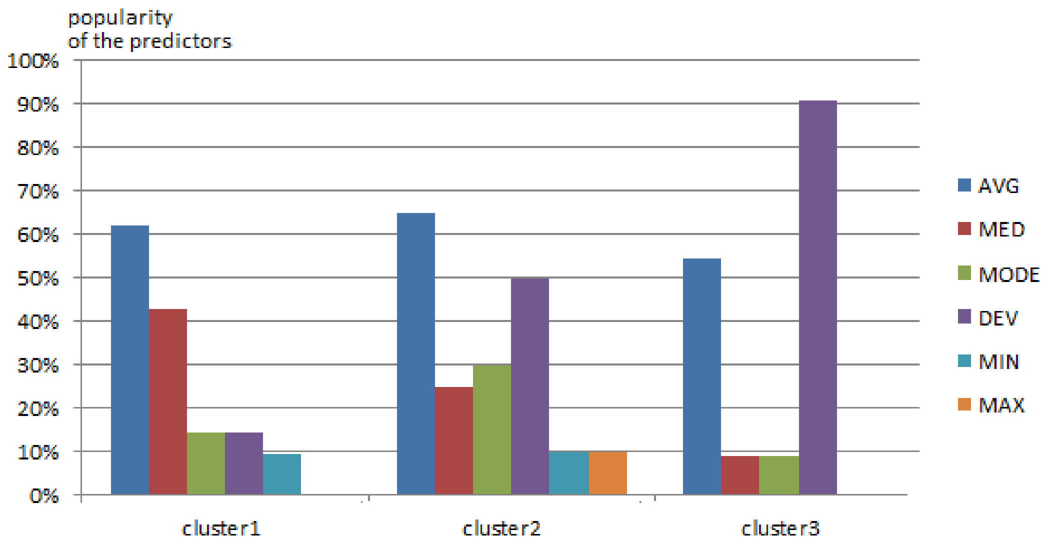



Figure 4.25: Populatiry of the predictors in individual formulas of users per cluster

# Chapter 5

# Data Organization and Management

For building the data structure capable of fast and convenient data access and operation, the data organization of the application dedicated for the same purpose and described in [11] was analyzed. Some ideas about the overall organization of tables and relationships between them were borrowed from this application, and some aspects required reconsideration. To induce the points that needed the most attention, the possible actions of the user were defined.

## 5.1 Motivation

For this project it is not important in what way the stress level was measured and how the data was collected. We base on the assumption that the stress level data is somehow obtained and needs to be stored and later visualized. In this section, we perform some experiments with real stress level data, so we describe the device used for the stress level measurement in this experiment, and the structure of the data it produces.

### 5.1.1 Device Description

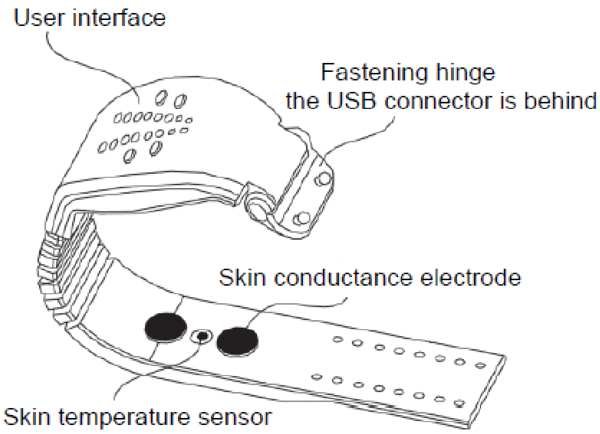
One of the recent researches dedicated to experiencing stress at work was held by Philips Research and Eindhoven University of Technology (TU/e). In this study, stress level measurement was based on inspecting sweat gland activity, reflected in skin conductivity. The actual measurements were performed by a device developed by Philips Research: *The Discrete Tension Indicator (DTI-2)*, which is an unobtrusive, wearable device that combines multiple sensors measuring skin conductance, 3D acceleration, band temperature, skin temperature and ambient light. Skin conductance, as measured by DTI-2, turns out to reflect stress reactions well enough to use it as basis for estimations of the stress levels of a person [22]. The skin conductance signal is sampled at the nodes at 10 Hz.

Figure 5.1 gives an idea of how the device looks and how it is worn.

The list below represents typical operations with data that can be performed by user:

- transferring data from the device to the application;
- re-filtering the signal;
- building the visualization for the selected period;
- selecting a part of the signal and zooming it in in order to see small details.

Based on the mentioned user actions, the data storage system should be designed.




a) overview of the device (from [12])

b) the device worn (from [12])

Figure 5.1: The Discrete Tension Indicator (DTI-2)

## 5.2 Data Organization Alternatives

The storing method should primarily satisfy the requirement of fast access to measurement data. Some user actions (e.g., building visualizations) will require processed data (filtered, smoothed and aggregated per minute values), others (e.g., viewing some parts of the signal in detail) will operate only with the raw data in the form it was obtained from the sensor. The operations of applying of a new filtering algorithm and transferring data from the device into the application work with both kinds of data. Therefore, both kinds of data should be necessarily stored somewhere. For each data kind, several storage alternatives can be proposed.

### 5.2.1 Raw Data

Raw data is the data in the same format as obtained from DTI-2. This kind of data is not supposed to be used for regular operations of data search or visualizations building. However, keeping raw data is necessary for research purpose, such as applying a new filtering algorithm, looking at the visualizations details at a bigger scale. Raw measurements can be stored in the following ways:

(1) Storing raw data in the database table.

The data recorded by DTI-2 is initially stored in files that have a unified structure, but different size, number of records and time of start and finish of recording (some days can be even spread between two or more files). At some points (e.g., because of the bad connection between skin and device or battery discharge), some parts of data can be missed. Copying all data from files to the database makes searching of necessary data very easy. On the other hand, recording data with the frequency of 10 Hz makes the final dataset for one user really huge (around 10 000 000 of records monthly for only working hours). The execution time of the queries to the database grows with the growth of the database. Too big database brings the disadvantage of time-consuming data search.

Another important question concerns the organization of the measurement data in a table (or several tables). The following possibilities can be considered.

(1) The measurement data of all users is stored in a single table.

This approach is a standard one, allowing building the queries simply and clearly. The possible bottleneck in this method is the amount of data necessary to keep in this table:

the dataset is going to be huge, making it impossible to operate with the database at the desired speed.

(2) A separate table is created for the measurements of each user.

This method has the advantage of quicker operation with data. The drawbacks of this approach are a larger number of tables, having more complicated names, harder maintainable database structure.

Assume that some desktop application is used by 10 people, and contains the data measured with 10 Hz frequency for the period of one year. In this case, the table with measurements will have at least 691 200 000 records (considering that the measurements are only taken during the working hours). As we see, a table containing the measurements of all users is enormously big, so it will be impossible to operate fast with it. Hence, the obvious choice is to store the data of each user in a separate table.

Concerning the devices worn by users, the following assumptions were made:

- one user can wear several devices, at the same time or sequentially;
- one device can be worn by different users at different moments of time.

Taking everything said above into account, we propose the following database organization. There should be a separate table with raw measurements for each pair "user-device". If a user gets a new device, a new table should be added into the database. Table "Device" should store information about the start and end time of wearing a device by a user. Figure 5.2 represents a fragment of the database schema related to the users and measurements.

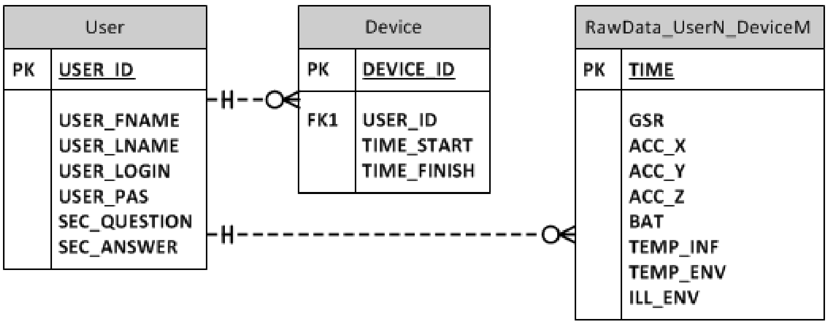



Figure 5.2: The fragment of the database schema related to the users and raw measurements

(2) Storing raw data in files reorganized in the way that each file contains exactly one day.

Taking into consideration that all files can have different amount of data and can include more than one day, or one day can be spread between two or more files, looking for a piece of data corresponding to a certain time period requires a complicated algorithm, which makes the data search time-consuming. Reorganization of the files, so that one file corresponds to exactly one day, seems to partially solve this problem. The major disadvantage of this approach is necessity to accomplish costly operations of reading files (e.g., when searching for a particular signal piece that is necessary to see in detail) and writing files (e.g., updating a file when the data is re-filtered).

The files organization in this case should be the following. The root folder called "*Raw*" should contain separate folders for each user named in the format "*UserN*", where $N$ is the personal number of the user. Inside this folder, there should be one or more folders named in the format "*DeviceM*", where $M$ is the number of the device worn by the user. Each folder corresponding to a particular user and a particular device should contain all raw data

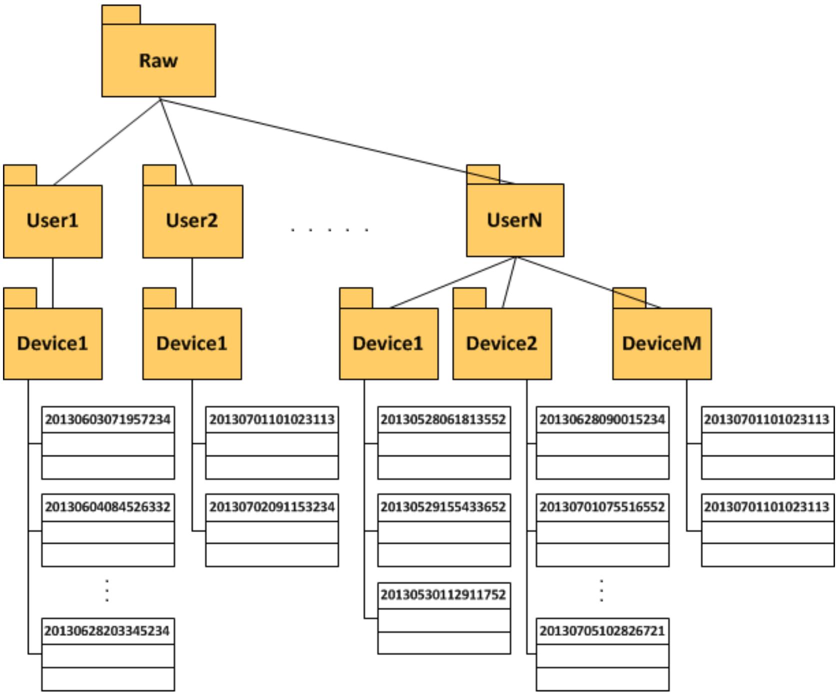



Figure 5.3: The folders and files organization for raw data

files, named in the format "*yyyyMMddHHmmssSSS*", where *yyyy* is the year, *MM* is the month, *dd* is the date, *HH* is the hour, *mm* is the minute, *ss* is the second and *SSS* is the millisecond of the first timestamp of the file. Figure 5.3 schematically represents the folders and files organization.

In addition to the file structure, a database indicating possible gaps in data within the files can be created. Storing the timestamps of the start and end of missing periods can be helpful in sense of a faster operation.

### 5.2.2 Processed Data

Processed data is filtered and aggregated per minute values. This data has lower density (frequency), and therefore takes less memory than the raw data. This data is used for building visualizations. For the processed data, we propose the following ways of storing the information.

(1) Storing processed data in the database table.

This method provides easy access to the processed data using database queries.

Taking into consideration the same assumptions as for the raw data, the following database organization is proposed. Since each user can wear several devices, there should be a separate table with processed measurements for each pair "user-device". If a user gets a new device, a new table should be added into the database. Table "Device" should store information about the start and end time of wearing a device by a user. Figure 5.4 represents a fragment of the database schema related to the users and measurements.

(2) Storing processed data in files reorganized in the way that each file contains exactly one day, besides, the filtered out values in the process data should be marked in order to differentiate between them and the missing data.

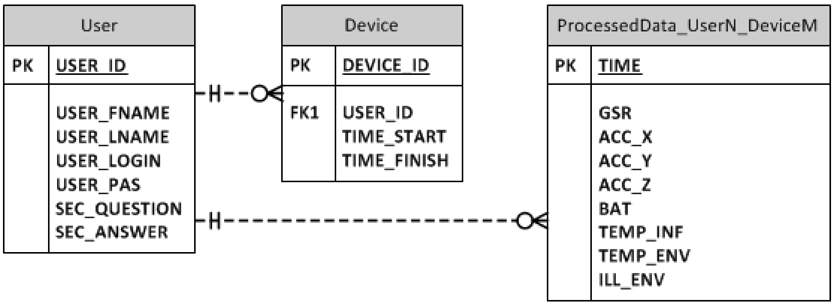



Figure 5.4: The fragment of the database schema related to the users and processed measurements

The filtered out values (unrealistically low or high numbers that are removed from the raw data as a result of filtering algorithm) should be marked with a certain label in order to differentiate between them and the empty values, when the device was not worn during the day. This way of storing data does not require creation of any database tables for the measurement data. In this alternative, the processed data files should be reorganized in the way that one file corresponds to exactly one day, and the files after the reorganization should be named using the corresponding date and user name for access simplicity. Files reorganization will make easier searching of the data necessary for visualizations building, which always requires the whole day.

The files organization in this case should be the following. The root folder called "*Processed*" should contain separate folders for each user named in the format "$UserN$", where $N$ is the personal number of the user. Inside this folder, there should be one or more folders named in the format "$DeviceM$", where $M$ is the number of the device worn by the user. Each folder corresponding to a particular user and particular device, should contain all raw data files, named in the format "$yyyyMMddHHmm$", where $yyyy$ is the year, $MM$ is the month, $dd$ is the date, $HH$ is the hour, $mm$ is the minute of the first timestamp of the file. Figure 5.5 schematically represents the folders and files organization.

In addition to the file structure, a database indicating possible gaps in data within the files can be created. Storing the timestapms of the start and end of missing periods can be helpful in sense of a faster operation.

(3) Storing processed data in files reorganized in the way that each file contains exactly one day, and does not have breaks if some values in the middle of the day are missing.

This alternative is almost the same as the alternative (2) with the difference that if some records in the middle of the day are missing (e.g., the device was not worn or was switched off), they are created manually for every missing minute with the value equals to $-1$. This is done in order to unify the file structure. This will simplify data search inside one file: it can be loaded in the memory, and since the time of the first timestamp and the frequency are known, it will be easy and fast to find the particular piece of data. The files organization in this case should be the same as in the alternative (2). This alternative may be useful for certain visualization types which require filled in with some values (e.g., by linear interpolation) gaps in data.

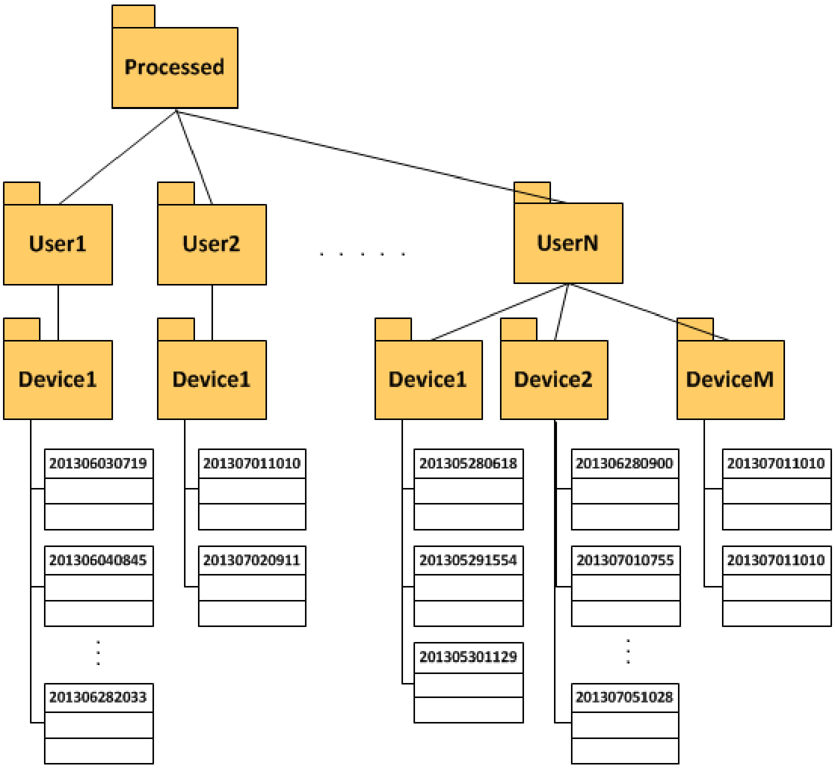



Figure 5.5: The folders and files organization for processed data

## 5.3 Evaluation

Among the mentioned alternatives of data organization, one should be taken. For each data organization alternative, some experiments were performed in order to estimate the possible execution time for the typical operations defined in section 5.1. For some experiments, some old data samples were used; for others, new data files were created manually by combining parts of real files, in order to imitate problems that can appear in real data files (e.g., large files on the edge between the days, multiple files corresponding to one day).

### 5.3.1 Datasets

For the first experiment, modeling the process of data transferring from the device to the application, we propose three datasets. Since users are supposed to monitor their stress data regularly, we assume that in most cases they transfer collected data from the device into the application daily. We also suppose that the interval between two transfers may vary from one day to two workweeks, and the same can be said about the amount of the data to be transferred. One of the tested datasets includes one file of 24 hours of measurements on the line between two days, another one contains the measurement of one workweek (around 62 hours of data), and the third one includes the measurements of two workweeks (124 hours). For the 24 hours dataset, a real data file generated by DTI-2 was taken. This file has the biggest possible size that so far can be generated by the device. Actually, it is a very unlikely situation that a user will wear the device continuously for almost 24 hours, but here testing of the critical file size took place. Two other datasets include the measurements from 5 (10) days spread among 8 (16) files. These measurements were generated manually following the structure of the real data, and they have more or less natural size: measuring stopped from time to time and continued later, modeling the process of switching the device off or not wearing it for some time. All three datasets are described in table 5.1.

Table 5.1: Datasets for the experiment #1.

| | 24 hours | 1 workweek | 2 workweeks |
|---|---|---|---|
| Number of files | 1 | 8 | 16 |
| Number of days | 2 | 5 | 10 |
| Number of hours | ≈24 | ≈62 | ≈124 |
| Size, MB | 74 | 189 | 379 |
| Frequency, Hz | 10 | 10 | 10 |
| Number of records | 856 147 | 2 254 310 | 4 508 620 |
| Number of processed records | 1 234 | 3 769 | 7 538 |

For the second experiment of selecting data for building visualizations, we have chosen four sets of processed data. Since visualizations can be built for different periods of time (from one day to one month), for testing the performance of the data structures the following datasets were chosen: 1 day, 1, 2 and 3 workweeks. The datasets are described in table 5.2.

Table 5.2: Datasets for the Experiment #2.

| | 1 day | 1 week | 2 weeks | 3 weeks |
|---|---|---|---|---|
| Number of files | 1 | 5 | 10 | 16 |
| Number of days | 1 | 5 | 10 | 16 |
| Number of hours | ≈17 | ≈62 | ≈124 | ≈186 |
| Size, KB | 16 | 73 | 142 | 215 |
| Number of records | 781 | 3 769 | 7 538 | 11 307 |

### 5.3.2 Conditions

All experiments were performed on the same computer in the same conditions, without any programs running in the background. Each experiment was performed 30 times registering the running time of each trial, whereupon the estimation of the expected execution time was calculated. Tables with the results of experiments present expected running times. For each experiment, a confidence interval for the estimation of expected execution time was built using t-statistics.

### 5.3.3 Experiments

**Experiment #1. Modelling the Operation of Transferring Data From the Device to the Application.**

This experiment models the process of data transfer from the device to the application. In real conditions this procedure includes the following steps:

(1) reading and copying to the memory the files with raw data (as they are obtained from DTI-2);

(2) filtering raw data;

(3) saving raw data to the database/files;

(4) downsampling data and saving it to the database/files.

Filtering algorithm is out of scope of the current project. Since the execution time of this step does not depend on the structure where the data is stored, it is not included in this experiment, though downsampling of the raw data is performed in order to create a proper data structure. Each step is performed for all proposed storage structures. To simulate natural conditions, the database tested was filled with four weeks of data before the experiment was performed.

We divide this experiment into two smaller ones, working with raw and processed data respectively.

**Experiment #1a**
This experiment includes working with raw data only, performing steps (1) and (3) described above.

### Input data

For running this experiment, only raw data is needed, which is taken in the original form of files as obtained from DTI-2. We run our experiment on two datasets, described in table 5.1. The database used in the experiment is preliminary filled in with 4 weeks of data.

### Output data

- for the database: filled in with data table structures described in figure 5.2.
- for files: files with data organized in a way described in figure 5.3.

### Expected execution times

The evaluation of expected execution times for this experiment are presented in table 5.3. Figure 5.6 shows the comparison of the confidence intervals for the expected values of the execution times for each dataset for the database and files alternatives when working with raw data. For each distribution, the standard deviation is specified.

Table 5.3: Experiment #1a. Reading files with raw data, reorganizing them and saving raw data to the database/files.

| Raw data / Datasets | Storage way DB | Files |
|---|---|---|
| 24 hours | 630 | 18 |
| 62 hours (1 workweek) | 1 816 | 46 |
| 124 hours (2 workweeks) | 3 838 | 92 |

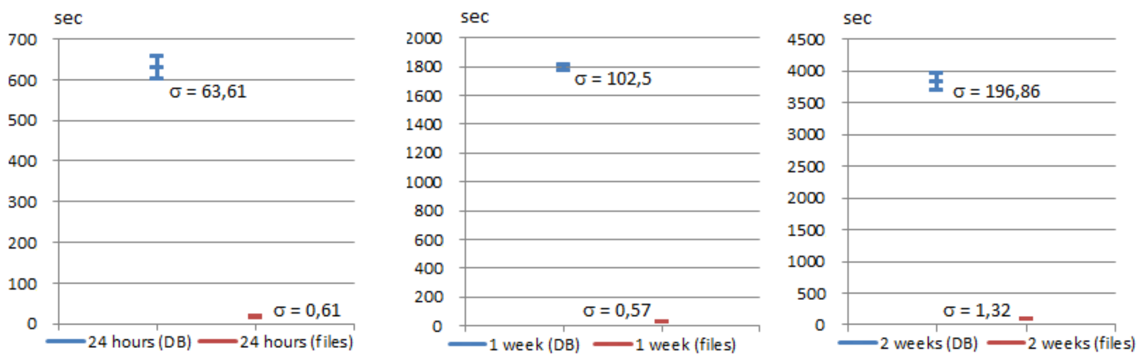



Figure 5.6: Experiment #1a. 95% confidence intervals for the database and files alternatives for raw data

**Experiment #1b**
This experiment works with processed data, including steps (1) and (4) described above.

**Input data**

The input data is the same as for experiment #1a: the datasets, described in table 5.1.

**Output data**

- for the database: filled in with data table structures described in figure 5.4.
- for files: files with data organized in a way described in figure 5.5.

**Expected execution times**

The evaluation of expected execution times for this experiment are presented in table 5.4. Figure 5.7 shows the comparison of the confidence intervals for the expected values of the execution times for each dataset for the database and files alternatives when working with processed data. For each distribution, the standard deviation is specified.

Table 5.4: Experiment #1b. Reading files with raw data, reorganizing and downsampling them and saving processed data to the database/files.

| Processed data / Datasets | Storage way: DB | Files | Files without gaps |
|---|---|---|---|
| 24 hours | 11, 8 | 10, 3 | 10, 7 |
| 62 hours (1 workweek) | 29, 3 | 28, 7 | 27, 9 |
| 124 hours (2 workweeks) | 57, 3 | 54, 3 | 55, 8 |

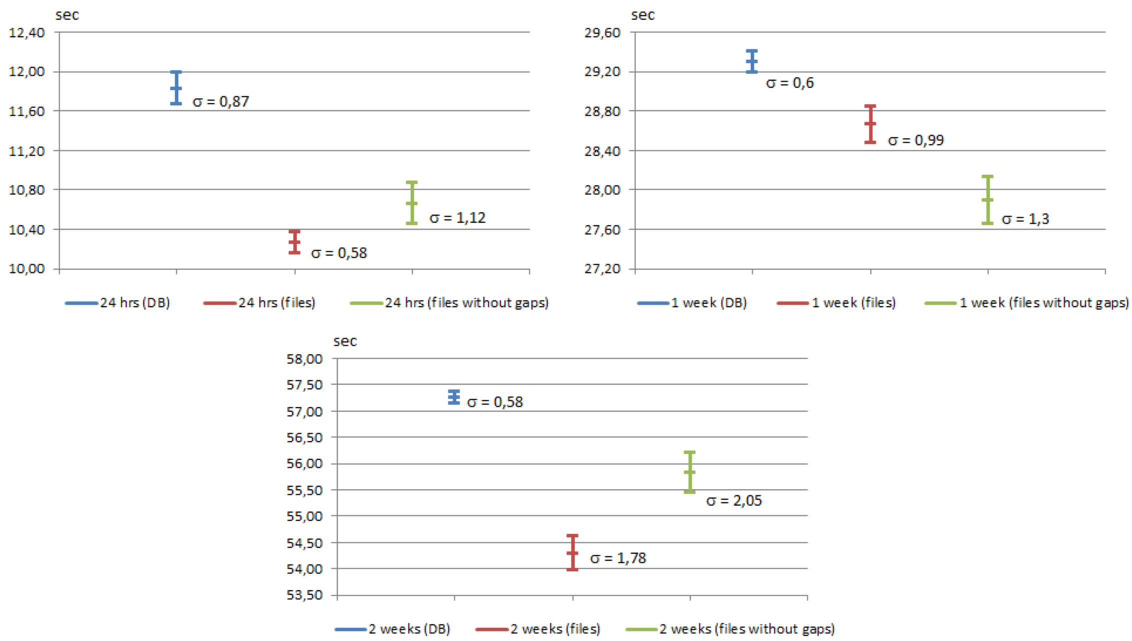



Figure 5.7: Experiment #1b. 95% confidence intervals for the database and files alternatives for processed data

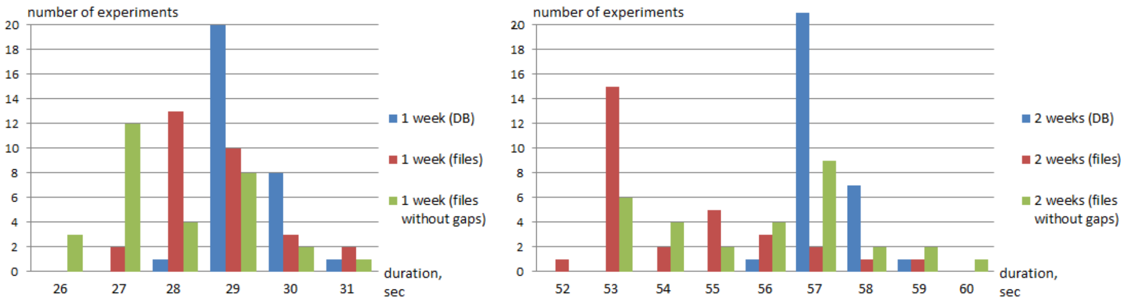



Figure 5.8: Experiment #1b. Histogram of experiment durations for the database and files alternatives for processed data

**Analysis**

Based on table 5.3, we can make a conclusion that storing raw data in the database is not an option, since the time necessary to add to the database even a small amount fo data (24 hours) is too long. Keeping raw data in files shows much better performance. Moreover, from figure 5.6 it can be seen that the standard deviation of the time needed to add datasets to the database is really huge. Such big differences in time can be explained by switching on and off the garbage collector of Java at random moments. As for the processed data, there is no big difference between the three proposed alternatives. Confidence intervals for the processed data shown in figure 5.7 do not have any intersections, though lie very close to each other. However, for processed data, normal files demonstrate the fastest performance for 24 hours and 2 weeks dataset, but for 1 week dataset the best performance is demonstrated by files without gaps. This fact might mean that normal files and files without gaps work with almost the same speed. Actually, the results demonstrated by all three kinds of data organization are really close and the differences between them are minimal, so we can consider that the performance of all of them is the same.

**Experiment #2. Modelling the Operation of Visualizations Building.**

The aim of this experiment is to estimate the time needed to select the necessary data amount from each of described data structures. At this stage, we only need to estimate and compare to each other the times needed for data selecting from different types of structures. Hence, during this experiment, we do not build any visualizations as they should look in the final application, as the time of building visualizations does not depend on the data structure chosen for the measurements storage, though we do create a simple line graph from the values selected in order to simulate the real conditions of the visualizations building.

**Input data**

For running this experiment, only processed data is necessary. We take the datasets, described in table 5.2. The data is now selected from the database table or files created during the Experiment #1:

- for the database: data from table "ProcessedData_UserM_DeviceN" described in figure 5.4.
- for files: data from the files organized in a way described in figure 5.5.

To get the necessary period of time from one of data structures, parameters in the form $[date1; date2]$ are used, where $date1$ is the start of the interval and $date2$ is its end.

**Output data**

The result of the experiment is one or more (depending on the number of days necessary to visualize) sets of data in a format suitable for visualizations building. The visualization type built in this experiment is a simple line graph, which uses the format called *XYDataSeries*, an interface that defines a collection of data in the form of $(x, y)$ values. The choice to use this interface was made because of the usage of one of the Java libraries (*JFreeChart*), which provides a wide range of opportunities for building various kinds of visualizations. The output data is the same for both kinds of input data.

### Expected execution times

The evaluation of expected execution times are presented in table 5.5. Figure 5.9 shows the comparison of the confidence intervals for the database and files alternatives for the execution times for each dataset. Since for none of the datasets the confidence intervals have intersection, we can say that the execution times for each dataset are different indeed.

### Analysis

The difference between selecting data from the database and files is not very significant for short periods of time, but becomes more noticeable along with the growth of the dataset. The bigger the dataset is, the better performance is demonstrated by files. Thus, for searching big amounts of data, the files are preferable to the database.

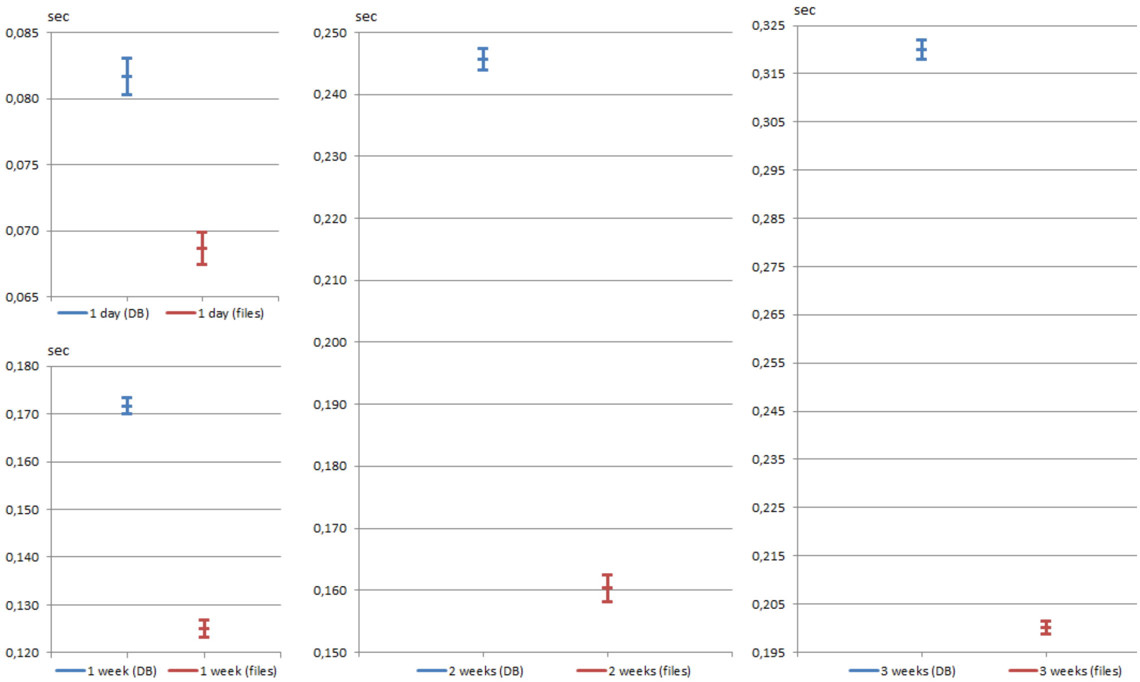



Figure 5.9: Experiment #2. 95% confidence intervals for the database and files alternatives

Table 5.5: Experiment #2. Selecting Data for Building Visualizations.

| Processed Data Stored | Execution Time, sec | |
|---|---|---|
| | DB | files |
| 1 day | 0, 08 | 0, 07 |
| 1 week | 0, 17 | 0, 13 |
| 2 weeks | 0, 25 | 0, 16 |
| 3 weeks | 0, 32 | 0, 20 |

## 5.4 Conclusion

The results of performance evaluation shown in tables 5.3, 5.4 and 5.5 allow to make a decision about the storage place for raw and processed data.

### Raw data

As it was already said above, raw data should be definitely stored in files (that is obvious from table 5.3 and figure 5.6).

### Processed data

For processed data, the situation is not so decisive. On the one hand, files (including files without gaps) demonstrate slightly better performance for both experiments (tables 5.4 and 5.5), though the difference in the execution times for experiments is only several seconds, which is also shown by the confidence intervals in figures 5.7 and 5.9. On the other hand, the execution times for the database demonstrate smaller deviation (figure 5.8), which means more stable performance of the database compared to files. Below the pros and cons for each data structure type are summarized.

**Database**

**Pros**

- simple data search using queries due to indexing;
- easy implementation;
- stable performance;
- easy maintenance.

**Cons**

- slightly slower performance (compared to any kind of files).

**Files**

**Pros**

- fast performance.

**Cons**

- harder data search (in case of necessity of more complicated queries);
- harder maintenance (compared to the database);
- harder implementation (compared to the database).

**Files without gaps**

**Pros**

- fast performance;
- ease of finding the gaps in data.

**Cons**

- harder data search (in case of necessity of more complicated queries);
- harder maintenance (compared to the database);
- harder implementation (compared to the database).

Hence, the final decision is to store raw data in files and processed data in the database.

# Chapter 6

# Conclusion and Future Work

This chapter concludes the paper. Section 6.1 presents the summary of the work done within the boundaries of current project. In section 6.2, possible directions for future research are highlighted.

## 6.1 Conclusion

In this work, we investigated stress data visualization techniques. An overview of related work on time-series data visualization, performed as the first step of the project, gave an insight for developing new ideas for visualizing stress data. Real stress data obtained from the sensor was represented by new visualization techniques. Three techniques from the list were validated in a small user study. This helped to estimate how vivid, clear and obvious each of the introduced visualizations is. In order to make one of the proposed visualizations more intuitive for users, another study was held in a small group of participants. The results of the experiments showed the following:

- user perception of the stress level changes depending on the visualization type shown. A picture of shape visualization gives the impression of higher stress level than a picture of color visualization built on the same data. The stress level on the colored shape visualization is perceived as the lowest;
- visualizations using colors are more intuitively understandable by users than the ones using form (shape). Users perform the best when working with color visualization rather than shape and colored shape ones;
- different visualization techniques make different tasks easier to perform: color shape visualization allows for easy detection of steep growth of the stress level, color one works best of all for finding the lowest values and small differences between the highest and the lowest values of the stress level;
- there is no significant difference in how male and female participants perceive visualizations. Slight variations in answers are only noticeable for separate days for the color and colored shape visualizations. We believe that these differences are caused by different color perception of men and women.
- even in the same visualization, different people pay attention to different factors. Clustering of the answers showed presence of several groups of users, each having its predictor, defining the answers of users in this group. However, the number of clusters is too big in comparison with the overall number of participants.

The initial goal to embed the proposed visualizations into one of existing calendar applications was shifted to the new one after the database structure problem was encountered. It was decided to develop a new data structure allowing for fast and easy data access. Two possible data structures were proposed: database and file structure. Several experiments were performed to estimate the speed of work of each alternative. The results of the performance evaluation showed the following:

- the data at high frequency of measurements can be stored only in files;
- the database works quite fast with the data at manually reduced frequency of measurements;
- in general, there is no significant improvement of the performance in comparison with the existing application, which makes senseless implementation of the new data structure.

Valuable and interesting results, both positive and negative, obtained in the project give inspiration for further research.

## 6.2 Future Work

Time limitations forced to strictly bound the project scope, leaving beyond the boundaries some ideas worth investigating. The following directions for the future work can be considered:

- continuation of the investigation on the visualization technique embedded into the month view of the calendar.

  The continuation of this study makes sense only with a bigger number of participants, since one of the conclusions of the cluster analysis described in chapter 4 was too small difference between the clusters and too big number of clusters compared to the overall number of participants.

- validation of the visualization techniques developed.

  In this project, we only chose three visualization techniques for the user study. There is still plenty of things to discover about these and other visualizations from the list of proposed, e.g., visualizations clarity, understandability, visibility of a particular kind of information, etc.

- embedding the visualizations into the existing application based on the conclusions and the chosen data storage method.

# Bibliography


[1] *Multidimensional Data Visualization Tools*. Proceedings of the 23rd SAS Users Group International, 1998. 13

[2] Wolfgang Aigner, Silvia Miksch, Wolfgang Müller, Heidrun Schumann, and Christian Tominski. Visualizing Time-Oriented Data — A Systematic View. *Computers & Graphics*, (31), 2007. 7, 8

[3] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of Time-Oriented Data*. Springer-Verlag London Limited, 2011. 7

[4] Rita Borgo and Heike Jänicke. Evaluating the Impact of Task Demands and Block Resolution on the Effectiveness of Pixel-based Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):963–972, 2010. 13

[5] Cynthia Brewer and Mark Harrower. Colorbrewer 2.0 — Color Advice for Cartography. `http://www.colorbrewer2.org/`, Accessed Online in June 2013. 8, 9, 13, 16

[6] John V. Carlis and Joseph A. Konstan. Interactive Visualization of Serial Periodic Data. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*. ACM, 1998. 11, 16

[7] Wikipedia The Free Encyclopedia. Heat Map. `http://en.wikipedia.org/wiki/Heat_map`. 12

[8] Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013. 9, 10

[9] Susan Havre, Beth Hetzler, and Lucy Nowell. ThemeRiver: Visualizing Theme Changes over Time. In *Proceedings of the IEEE Symposium on Information Vizualization*. IEEE Computer Society, 2000. 10

[10] Tom Haynes. Email Visualization. `http://www.tom-haynes.com/email-vis#email-vis`, 2010. 11, 12

[11] Rafal Kocielnik, Fabrizio Maria Maggi, Natalia Sidorova, Joyce H.D.M. Westerink, and Martin Ouwerkerk. Smart Technologies for Long-Term Stress Monitoring at Work. 4, 5, 6, 43

[12] Koninklijke Philips Electronics N.V. *DTI-2. User Manual*, 2012. 44

[13] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. AffectAura: An Intelligent System for Emotional Memory. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 849–858. ACM, 2012. 4, 5

[14] Wolfgang Müller and Heidrun Schumann. Visualization Methods for Time-Dependent Data — an Overview. In *Proceedings of the 2003 Winter Simulation Conference*, 2003. 9, 10

[15] John Nelson. Five Years of Traffic Fatalities. http://uxblog.idvsolutions.com/2012/12/five-years-of-traffic-fatalities.html, 2012. 12

[16] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. Mind the Body! Designing a Mobile Stress Management Application Encouraging Personal Reflection. In *Proceedings of the 8th ACM conferenct on designing interactive systems*, pages 47–56. ACM, 2010. 6, 11

[17] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer Methods and Programs in Biomedicine*, 108(3):1287–1301, 2012. 4

[18] David Smith and Paul Bleicher. Charting time series as calendar heat maps in R. http://blog.revolutionanalytics.com/2009/11/charting-time-series-as-calendar-heat-maps-in-r.html, 2009. 12, 13

[19] Inc. Square. Cubism.js. Time Series Visualization. http://square.github.io/cubism/, Accessed Online in October 2013. 19

[20] Jarke J. van Wijk. The value of visualization. In C. Silva, E. Groeller, and H. Rushmeier, editors, *Proceedings of IEEE Visualization Conference*, pages 79–86, 2005. 2

[21] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004. 8

[22] Joyce H.D.M. Westerink, Martin Ouwerkerk, Gert-Jan de Vries, Stijn de Waele, Jack van den Eerenbeemd, and Marco van Boven. Emotion measurement platform for daily life situations. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009. 5, 6, 43

[23] Hugh R. Wilson and Frances Wilkinson. Symmetry Perception: a Novel Approach for Biological Shapes. *Vision Research*, 42(5):589–597, 2002. 18

[24] Nathan Yau. Five Years of Traffic Fatalities. http://flowingdata.com/2013/01/08/five-years-of-traffic-fatalities/, 2012. 12

# Appendix A

# Tables and Figures

Table A.1: Kruskal-Wallis test

| week | day | all visualization types | pairwise comparison | | |
|---|---|---|---|---|---|
| | | | color and colored shape | color and shape | shape and colored shape |
| 1 | Tuesday | 0,170 | 0,125 | 0,874 | 0,086 |
| | Wednesday | 0,000 | 0,014 | 0,002 | 0,000 |
| | Thursday | 0,005 | 0,256 | 0,021 | 0,002 |
| | Friday | 0,000 | 0,054 | 0,016 | 0,000 |
| 2 | Monday | 0,000 | 0,000 | 0,036 | 0,000 |
| | Tuesday | 0,000 | 0,074 | 0,021 | 0,000 |
| | Wednesday | 0,000 | 0,059 | 0,000 | 0,000 |
| | Thursday | 0,000 | 0,023 | 0,010 | 0,000 |
| | Friday | 0,011 | 0,188 | 0,125 | 0,002 |
| 3 | Monday | 0,000 | 0,141 | 0,000 | 0,000 |
| | Tuesday | 0,000 | 0,000 | 0,014 | 0,000 |
| | Wednesday | 0,000 | 0,000 | 0,161 | 0,000 |
| | Thursday | 0,013 | 0,004 | 0,976 | 0,029 |
| | Friday | 0,133 | 0,201 | 0,304 | 0,065 |
| 4 | Monday | 0,008 | 0,929 | 0,009 | 0,007 |
| | Tuesday | 0,000 | 0,857 | 0,002 | 0,000 |
| | Wednesday | 0,000 | 0,400 | 0,002 | 0,000 |
| | Thursday | 0,001 | 0,084 | 0,011 | 0,000 |
| | Friday | 0,000 | 0,073 | 0,013 | 0,000 |
| 5 | Monday | 0,005 | 0,379 | 0,030 | 0,001 |
| | Tuesday | 0,000 | 0,086 | 0,000 | 0,000 |
| | Wednesday | 0,000 | 0,007 | 0,004 | 0,000 |
| | Thursday | 0,000 | 0,110 | 0,000 | 0,000 |
| | Friday | 0,000 | 0,001 | 0,113 | 0,001 |
| 6 | Monday | 0,319 | 0,097 | 0,495 | 0,590 |
| | Tuesday | 0,000 | 0,000 | 0,175 | 0,000 |
| | Wednesday | 0,000 | 0,000 | 0,034 | 0,000 |
| | Thursday | 0,000 | 0,001 | 0,000 | 0,000 |
| | Friday | 0,000 | 0,000 | 0,395 | 0,000 |

Table A.2: Ranks (Kruskal-Wallis test)

| week | day | color | shape | colored shape |
|---|---|---|---|---|
| 1 | Tuesday | 48, 50 | 49, 67 | 38, 37 |
| | Wednesday | 52, 15 | 74, 18 | 35, 32 |
| | Thursday | 47, 99 | 63, 11 | 41, 15 |
| | Friday | 52, 54 | 70, 13 | 38, 75 |
| 2 | Monday | 57, 51 | 71, 04 | 32, 92 |
| | Tuesday | 52, 46 | 68, 75 | 40, 14 |
| | Wednesday | 48, 57 | 76, 44 | 36, 76 |
| | Thursday | 53, 08 | 71, 57 | 36, 85 |
| | Friday | 52, 61 | 64, 28 | 42, 72 |
| 3 | Monday | 38, 51 | 76, 25 | 47, 00 |
| | Tuesday | 59, 69 | 70, 06 | 31, 67 |
| | Wednesday | 58, 65 | 66, 24 | 34, 33 |
| | Thursday | 60, 29 | 58, 90 | 41, 61 |
| | Friday | 52, 18 | 58, 41 | 44, 33 |
| 4 | Monday | 48, 06 | 66, 89 | 47, 42 |
| | Tuesday | 47, 06 | 70, 79 | 44, 63 |
| | Wednesday | 48, 60 | 70, 49 | 43, 38 |
| | Thursday | 52, 17 | 68, 77 | 41, 47 |
| | Friday | 52, 40 | 68, 93 | 41, 08 |
| 5 | Monday | 50, 43 | 66, 47 | 43, 88 |
| | Tuesday | 49, 63 | 73, 26 | 39, 65 |
| | Wednesday | 52, 99 | 70, 57 | 36, 96 |
| | Thursday | 48, 81 | 73, 91 | 39, 83 |
| | Friday | 57, 79 | 65, 87 | 38, 67 |
| 6 | Monday | 56, 54 | 51, 06 | 46, 44 |
| | Tuesday | 62, 11 | 68, 82 | 30, 42 |
| | Wednesday | 58, 18 | 70, 74 | 32, 54 |
| | Thursday | 52, 33 | 73, 46 | 33, 81 |
| | Friday | 61, 69 | 64, 81 | 34, 63 |

Table A.3: 3-Fold Cross-Validation Results for the Color Visualization (Median answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $1,416 \cdot AVG_{color} - 1,433$ | $0,206$ |
| $round_2$ | $1,345 \cdot AVG_{color} - 1,133$ | $0,184$ |
| | $0,788 \cdot DEV_{color} + 1,081 \cdot AVG_{color} - 1,055$ | $0,207$ |
| | $\boldsymbol{0,616 \cdot MED_{color} + 0,872 \cdot DEV_{color} + 0,390 \cdot AVG_{color} - 0,889}$ | **0,181** |
| | $0,908 \cdot MED_{color} + 0,996 \cdot DEV_{color} - 0,653$ | $0,190$ |
| $round_3$ | $1,365 \cdot AVG_{color} - 1,066$ | $0,226$ |

Table A.4: 3-Fold Cross-Validation Results for the Shape Visualization (Median answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $0,051 \cdot AVG_{value} + 1,651$ | $0,225$ |
| | $-0,036 \cdot MIN_{value} + 0,050 \cdot AVG_{value} + 2,019$ | $0,289$ |
| $round_2$ | $\boldsymbol{0,053 \cdot AVG_{value} + 1,629}$ | **0,153** |
| $round_3$ | $0,055 \cdot AVG_{value} + 1,629$ | $0,346$ |

Table A.5: 3-Fold Cross-Validation Results for the Colored Shape Visualization (Median answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $0,051 \cdot AVG_{value} + 0,036$ | $0,215$ |
| | $0,034 \cdot DEV_{value} + 0,035 \cdot AVG_{value} + 0,055$ | $0,154$ |
| $round_2$ | $0,044 \cdot AVG_{value} + 0,611$ | $0,311$ |
| | $0,803 \cdot DEV_{color} + 0,033 \cdot AVG_{value} + 0,357$ | $0,222$ |
| $round_3$ | $0,052 \cdot AVG_{value} + 0,228$ | $0,196$ |
| | $\boldsymbol{0,035 \cdot DEV_{value} + 0,035 \cdot AVG_{value} + 0,226}$ | **0,130** |

Table A.6: 3-Fold Cross-Validation Results for the Color Visualization (Average answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $1,368 \cdot AVG_{color} - 1,205$ | $0,212$ |
| | $0,714 \cdot DEV_{color} + 1,054 \cdot AVG_{color} - 0,889$ | $0,120$ |
| $round_2$ | $1,273 \cdot AVG_{color} - 0,748$ | $0,179$ |
| | $1,018 \cdot DEV_{color} + 0,932 \cdot AVG_{color} - 0,648$ | $0,174$ |
| | $-0,191 \cdot MIN_{color} + 0,811 \cdot DEV_{color} + 1,007 \cdot AVG_{color} - 0,416$ | $0,166$ |
| $round_3$ | $1,268 \cdot AVG_{color} - 0,619$ | $0,158$ |
| | $\boldsymbol{0,795 \cdot DEV_{color} + 0,928 \cdot AVG_{color} - 0,397}$ | **0,100** |
| | $0,374 \cdot MED_{color} + 0,892 \cdot DEV_{color} + 0,510 \cdot AVG_{color} - 0,305$ | $0,101$ |

Table A.7: 3-Fold Cross-Validation Results for the Shape Visualization (Average answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $0,044 \cdot AVG_{value} + 2,374$ | $0,147$ |
| | $\boldsymbol{0,031 \cdot DEV_{value} + 0,030 \cdot AVG_{value} + 2,391}$ | **0,134** |
| | $-0,024 \cdot MIN_{value} + 0,027 \cdot DEV_{value} + 0,031 \cdot AVG_{value} + 2,636$ | $0,201$ |
| $round_2$ | $0,045 \cdot AVG_{value} + 2,249$ | $0,185$ |
| | $-0,021 \cdot MIN_{value} + 0,047 \cdot AVG_{value} + 2,383$ | $0,186$ |
| $round_3$ | $0,046 \cdot AVG_{value} + 2,329$ | $0,257$ |
| | $0,022 \cdot DEV_{value} + 0,036 \cdot AVG_{value} + 2,327$ | $0,227$ |

Table A.8: 3-Fold Cross-Validation Results for the Colored Shape Visualization (Average answer)

| | model formula | error |
|---|---|---|
| $round_1$ | $0,048 \cdot AVG_{value} + 0,413$ | $0,184$ |
| | $\boldsymbol{0,711 \cdot DEV_{color} + 0,036 \cdot AVG_{value} + 0,307}$ | **0,131** |
| $round_2$ | $0,045 \cdot AVG_{value} + 0,697$ | $0,192$ |
| | $0,824 \cdot DEV_{color} + 0,034 \cdot AVG_{value} + 0,436$ | $0,145$ |
| $round_3$ | $0,048 \cdot AVG_{value} + 0,659$ | $0,192$ |
| | $0,026 \cdot DEV_{value} + 0,035 \cdot AVG_{value} + 0,657$ | $0,133$ |

Table A.9: 3-Fold Cross-Validation Results for Clusters

| cluster | round | model formula | error |
|---|---|---|---|
| $cluster_1$ | $round_1$ | $1,231 \cdot AVG_{color} - 0,728$ | $1,828$ |
| | $round_2$ | $0,953 \cdot MED_{color} + 0,467$ | $2,283$ |
| | | $-0,279 \cdot MIN_{color} + 0,959 \cdot MED_{color} + 0,844$ | $2,064$ |
| | | $0,534 \cdot AVG_{color} - 0,285 \cdot MIN_{color} + 0,515 \cdot MED_{color} + 0,498$ | $1,718$ |
| | $round_3$ | $1,002 \cdot MED_{color} + 0,605$ | $2,390$ |
| | | $0,431 \cdot DEV_{color} + 0,865 \cdot MED_{color} + 0,502$ | $1,879$ |
| $cluster_2$ | $round_1$ | $1,616 \cdot AVG_{color} - 2,261$ | $3,020$ |
| | | $0,823 \cdot DEV_{color} + 1,255 \cdot AVG_{color} - 1,897$ | $2,022$ |
| | | $0,977 \cdot MED_{color} + 1,432 \cdot DEV_{color} - 0,152 \cdot AVG_{color} - 0,840$ | $2,054$ |
| | | $0,878 \cdot MED_{color} + 1,358 \cdot DEV_{color} - 0,972$ | $1,939$ |
| | $round_2$ | $1,469 \cdot AVG_{color} - 1,555$ | $2,565$ |
| | | $1,237 \cdot DEV_{color} + 1,055 \cdot AVG_{color} - 1,434$ | $2,147$ |
| | | $-0,266 \cdot MAX_{color} + 1,974 \cdot DEV_{color} + 1,134 \cdot AVG_{color} - 0,786$ | $2,263$ |
| | $round_3$ | $1,572 \cdot AVG_{color} - 1,782$ | $2,142$ |
| | | $1,061 \cdot DEV_{color} + 1,117 \cdot AVG_{color} - 1,486$ | $1,549$ |
| | | $0,633 \cdot MED_{color} + 1,225 \cdot DEV_{color} + 0,411 \cdot AVG_{color} - 1,331$ | $1,520$ |
| | | $0,932 \cdot MED_{color} + 1,385 \cdot DEV_{color} - 1,060$ | $1,552$ |
| | | $-0,243 \cdot MAX_{color} + 0,979 \cdot MED_{color} + 2,018 \cdot DEV_{color} - 0,334$ | $1,584$ |
| $cluster_3$ | $round_1$ | $3,126 \cdot DEV_{color} + 0,866$ | $4,416$ |
| | | $0,400 \cdot MED_{color} + 2,639 \cdot DEV_{color} - 0,172$ | $3,561$ |
| | $round_2$ | $3,510 \cdot DEV_{color} + 0,503$ | $3,399$ |
| | | $1,007 \cdot AVG_{color} + 2,162 \cdot DEV_{color} - 2,009$ | $2,270$ |
| | $round_3$ | $2,984 \cdot DEV_{color} + 1,114$ | $3,076$ |
| | | $0,914 \cdot AVG_{color} + 2,006 \cdot DEV_{color} - 1,265$ | $2,500$ |

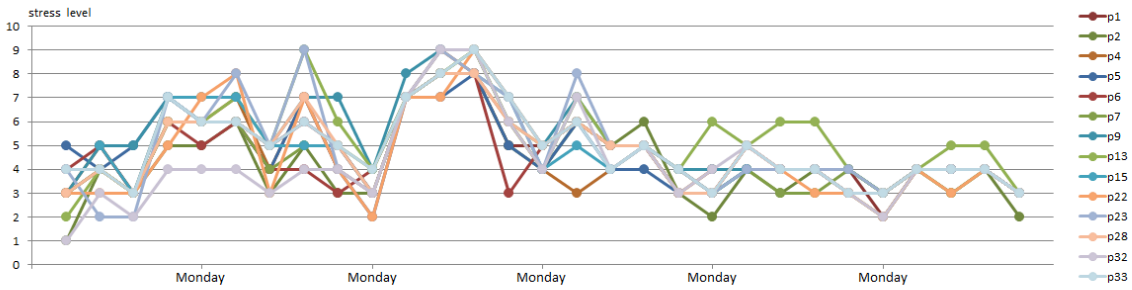



Figure A.1: Clustering analysis of participants. Cluster 1

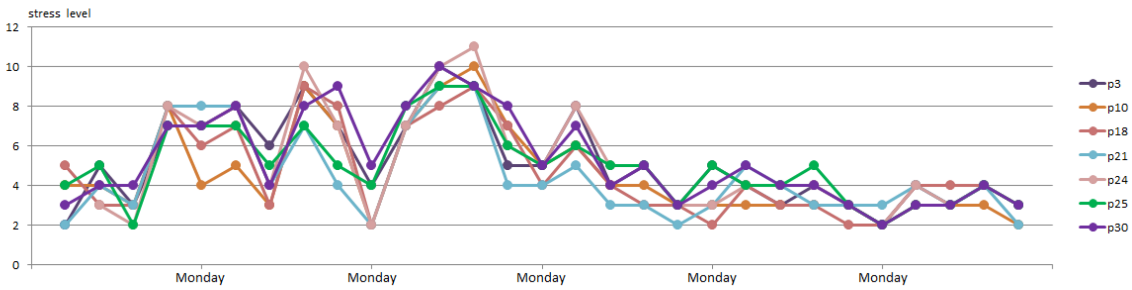



Figure A.2: Clustering analysis of participants. Cluster 2

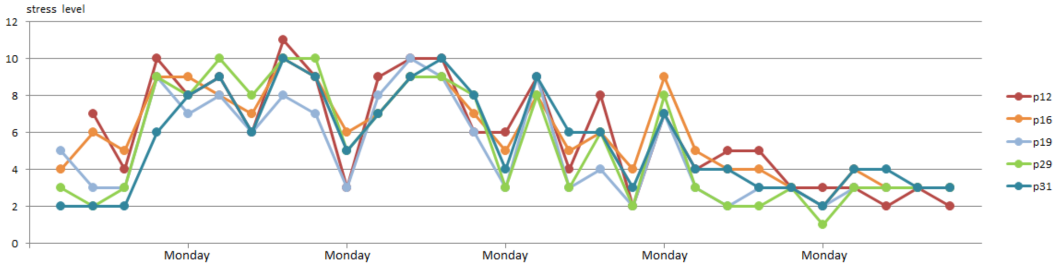



Figure A.3: Clustering analysis of participants. Cluster 3

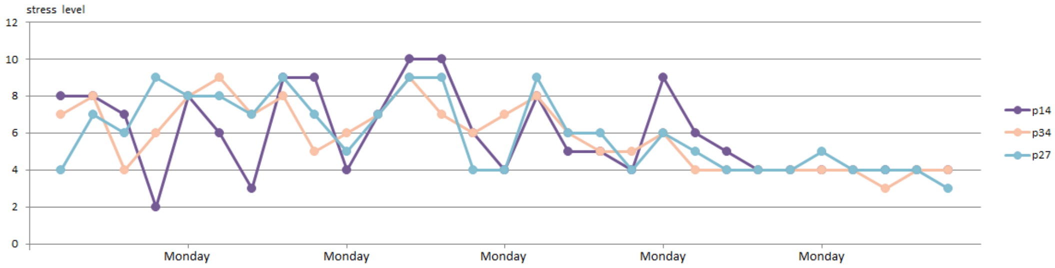



Figure A.4: Clustering analysis of participants. Cluster 4

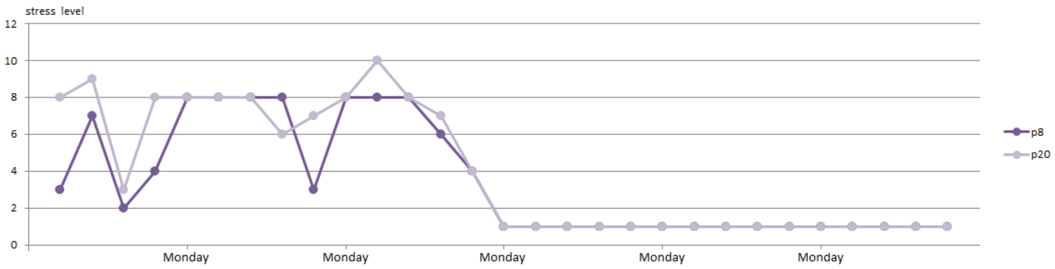



Figure A.5: Clustering analysis of participants. Cluster 5

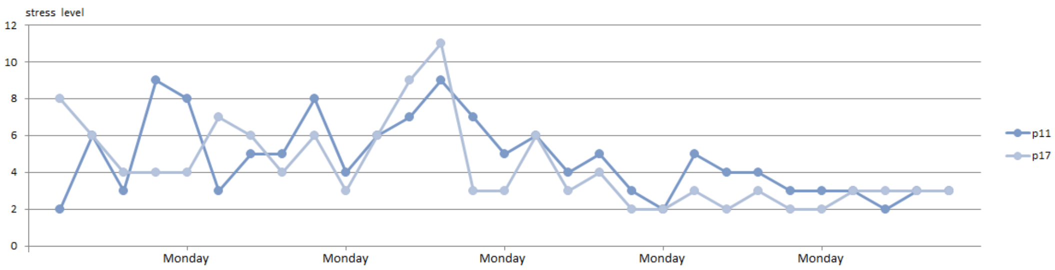



Figure A.6: Clustering analysis of participants. Cluster 6

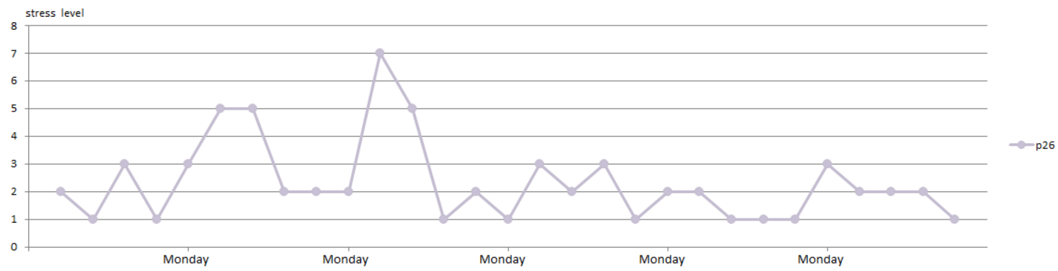



Figure A.7: Clustering analysis of participants. Cluster 7