DEPARTMENT OF ENGINEERING MATHEMATICS

Exploring Multilevel Monitoring for sleep quality and its associated factors

Pantelis Zoumpoulidis

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Friday 17th September, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Pantelis Zoumpoulidis, Friday 17th September, 2021

Contents

1 Introduction **1**
- 1.1 Topic Importance . . . 1
- 1.2 Dataset . . . 2
- 1.3 Objectives . . . 2
- 1.4 Challenges . . . 3
- 1.5 Summary . . . 3

2 Literature Review **5**
- 2.1 Feature Extraction from Time Series . . . 5
- 2.2 External Factors, Heart Rate, and Sleep Quality . . . 6
- 2.3 Previous work using the MMASH dataset . . . 7
- 2.4 Summary . . . 7

3 Methodology **9**
- 3.1 Importing the Data . . . 9
- 3.2 Preprocessing the Data . . . 9
- 3.3 Analyzing the Data . . . 11
- 3.4 Summary . . . 15

4 Results **19**
- 4.1 Statistical Terminology . . . 19
- 4.2 Heart Rate Clustering Results . . . 20
- 4.3 Associated Factors Clustering Results . . . 26
- 4.4 Heart Rate and Associated Factors Combination . . . 29
- 4.5 Summary . . . 32

5 Further Work **33**
- 5.1 Different Techniques . . . 33
- 5.2 Different Features . . . 34

6 Conclusions **35**
- 6.1 Heart Rate Patterns of Sleep Quality . . . 35
- 6.2 Associated Factors of Sleep Quality . . . 35
- 6.3 Combination of Heart Rate and Associated Factors . . . 36

A Clustering Results **43**

B Execution Instructions **45**

List of Figures

3.1 Standardized HR against Date & Time for Every User . . . 13

4.1 An example of sigmoid function . . . 20
4.2 Rand Score using extracted features with KMeans and AC . . . 21
4.3 Rand Score using selected extracted features with KMeans and AC . . . 21
4.4 Rand Score using TSLearn KMeans with DTW . . . 21
4.5 Rand Score using Motifs . . . 22
4.6 Motifs with AC . . . 23
4.7 Motifs with KMeans . . . 23
4.8 Rand Score using Discords . . . 24
4.9 Discords with MyClusterAlgorithm . . . 24
4.10 Rand score using Snippets . . . 25
4.11 Snippets with MyClusterAlgorithm . . . 25
4.12 Rand score using Questionnaire Raw Results . . . 26
4.13 Multinomial Logistic Regression's coefficients of ass. factors using questionnaire results with AC . . . 26
4.14 Rand score using Questionnaire Labels Results . . . 27
4.15 Multinomial Logistic Regression's coefficients of ass. factors using questionnaire labels with AC . . . 27
4.16 Rand score using Activity Statistics Results . . . 27
4.17 Multinomial Logistic Regression's coefficients of ass. factors using activity statistics with KMeans . . . 28
4.18 PCC of ass. factors using activity statistics with the actual PSQI . . . 28
4.19 Rand Score of combining sleep quality and associated factors clustering results . . . 29
4.20 Best combined Sleep Quality and Associated Factors methods and their predictions . . . 29
4.21 Motifs of 90s length clustered with MyClusterAlgorithm . . . 30
4.22 Parallel Coordinates plot of the last configuration for Associated Factors with the AC predictions . . . 31
4.23 PSQI - Cortisol after sleep . . . 32

A.1 Sleep Quality Clustering Results . . . 43
A.2 Associated Factors Clustering Results . . . 43

List of Tables

3.1 PSQI to PSQI Label 10
3.2 z-Normalized Euclidean Distances Matrix 12
3.3 Description of Associated Factors 16

4.1 Rand Index Calculation Table 19

Abstract

The technological revolution that started many years ago still continues nowadays. Devices that were big in size and technologies that were accessible only to big organizations are now compact and commercially available. Wearable devices, like smartwatches, are an example that can provide information about movement and heart rate, which once was only available in research labs or hospitals using special equipment. That created the opportunity for more scientists to do more researches easier and faster. One of the fields affected by that is sleep, which is vital for our health, performance, and wellbeing.

Many studies have been conducted to discover heart rate patterns with sleep quality; however, to the best of our knowledge, none of them used our methodology. Our approach was focused on extracting unique features from the time series, like motifs, discords, and shapelets. Also, the dataset we used was rich in terms of features, allowing us to investigate more thoroughly various factors and their association to sleep quality. We found that caffeinated drink consumption, sitting, and age are factors that negatively affect the quality of sleep. The opposite applies to slow/medium walking, height, and the number of steps. There were also indications that heavy training and negative feelings prior to bed time also affects its quality negatively. Last but not least, we have strong evidence that a couple of participants reported a different sleep quality than the one their data suggest. We strengthened our assumption by comparing the cortisol levels of our dataset with the prevailing theory.

- I spend 120 hours collecting material on and learning about the current state-of-the-art methodologies in time series clustering.
- I wrote a total of 2568 lines of source code in Python 3.9, implementing the algorithms discussed in this thesis.
- I used various tests to associate heart rate patterns with sleep quality.
- I conducted numerous amounts of experiments to determine how external factors affect sleep.

Supporting Technologies

- I used PyCharm to develop the Python 3.9 code and execute it.
- I used Microsoft Excel to display the clustering results.
- I used LaTeX to format the thesis, via the online service *Overleaf*.

Notation and Acronyms

AC	:	Agglomerative Clustering
AHBLL	:	Adaptive Heartbeat Locked Loop
AR	:	Activity Recognition
BIS/BAS	:	Behavioral Avoidance/Inhibition
DBA	:	Dtw Barycenter Averaging
DSI	:	Daily Stress Inventory
DTW	:	Dynamic Time Warping
ECG	:	Electrocardiogram
EEG	:	Electroencephalogram
FN	:	False Negative
FP	:	False Positive
HMM	:	Hidden Markov Models
HR	:	Heart Rate
HRV	:	Heart Rate Variability
MCFS	:	Multi-Cluster Feature Selection
MDL	:	Minimum Description Length
MEQ	:	Morningness-Eveningness
ML	:	Machine Learning
MMASH	:	Multilevel Monitoring of Activity and Sleep in Healthy People
PAA	:	Piecewise Aggregate Approximation
PANAS	:	Positive and Negative Affect
PCA	:	Principal Component Analysis
PCC	:	Pearson Correlation Coefficient
PSG	:	Polysomnography
PSQI	:	Pittsburgh Sleep Quality Index
RFS	:	Restricted Forward feature Selection
RI	:	Rand Index
rMSSD	:	root Mean Squared error of Successive Inter-Beat Intervals Differences
SAX	:	Symbolic Aggregate approXimation
SDNN	:	Standard Deviation of Inter-Beat Intervals
SDNN24	:	Standard Deviation of Inter-Beat Intervals over 24 hours
STAI-Y	:	State-Trait Anxiety Inventory
TN	:	True Negative
TP	:	True Positive
TSC	:	Time Series Classification
TST	:	Total Sleep Time
u-shapelets	:	Unsupervised-Shapelets
WASO	:	Wake After Sleep Onset
W-k-means	:	Weighted K-Means

Acknowledgements

I want to say a huge thanks to my father, Mr. Pavlos Zoumpoulidis, for supporting me both psychologically and financially for the duration of my studies.

Chapter 1

Introduction

This chapter begins by presenting the problem we investigated. The first section underlines why it is an important subject, and after that, a segment focuses on the dataset we used for our research. Following, we outline the initials aims of this thesis. Next, we present the challenges involved with this project and their significance, and finally, we present a summary of the chapter. Before we begin, however, we note that there is also a video presentation of this thesis available online [73].

Over the last decades, technology has seen an undeniable revolution in creating smart devices for every place on Earth, from a field on a mountain to the wrist of our hands. These innovative devices have given us the ability to obtain real-time information, which consequently created the capability to track and classify people's activities and behaviors in an automated way. Machine Learning (ML) [29] methods are adapted to accurately assess and track changes in physical activity and patterns measured by various activity sensors. Due to the important beneficial effects on physical and mental health, Activity Recognition (AR) [7] has been considered a fundamental paradigm for smart healthcare and wellness.

Nowadays, it is easier than ever to obtain information about an individual's sleep since many wearable devices can capture that kind of intelligence. That allowed more scientists to do more researches, as we show in chapter 2. Some examples of these devices are Apple Watch, Fitbit, ActiGraph, Jawbone, Pebble [59], Galaxy Wearable, and Zeo [12]. In order to collect physical activity and sleep data, they use one or a combination of technologies like accelerometers, gyroscopes, electrocardiogram (ECG), and electroencephalogram (EEG) sensors. Most of these devices store the data on the cloud, but all of them make it available to the user to download the data through an accompanying app or website and analyze them. The Heart Rate (HR) and accelerometer data are collected in a very high frequency, even multiple times each second. That produces a time series of n x m dimensions, where n is the number of different features and m are the different timestamps when the data points are recorded [59]. In general, time series is a set of observations made sequentially through time [63]. That is where AR becomes essential, as it can automatically identify activities in a time series, like walking or sitting.

1.1 Topic Importance

Every living organism requires sleep. It is vital for our health, performance, and wellbeing [53]. It is also highly related to some diseases [6] and a wide variety of health problems [12]. Ideally, sleep should consume about seven to eight hours of our everyday lives [21], almost one-third of our day. However, recent research has shown that most adults sleep on average 6.8 hours daily, 2.2 hours less than a century ago, with a third of them sleeping less than 6 hours [31]. On the same wavelength, people that sleep less than 6 hours a day are keener on associating with fair or poor health. Surprisingly, the same applies to individuals that sleep for more than 9 hours a night [32]. Similarly, Miwa et al. [44] mention that increased sleep durations have a negative impact on sleep quality because only light sleep increases toward the end of a sleep cycle. Hence, one can quickly apprehend how vital sleep, and consequently, this topic is.

Additionally, this thesis intends to benefit a wide variety of people with different backgrounds. Firstly, it can be valuable to individuals who try to figure out how to improve their sleep quality. Likewise, doctors or sleep experts can read this research to learn about our findings so they can have additional knowledge when they come across a similar circumstance. Moreover, developers can use our approach in their

applications to provide feedback to the users in order to improve their sleep quality. Finally, it can be used by other data scientists as they can apply this kind of methodology to different problems.

1.2 Dataset

In our research, we use the Multilevel Monitoring of Activity and Sleep in Healthy People (MMASH) dataset, published by Rossi et al. [55]. It consists of observations from 22 healthy young adult males (age = 27.29 ± 4.21 years; height = 179.91 ± 8.22 cm; weight = 75.05 ± 12.79 kg), with the majority of them being students at the University of Pisa. At the beginning of the study, the participant filled questionnaires about Morningness-Eveningness (MEQ), State-Trait Anxiety Inventory (STAI-Y), Pittsburgh Sleep Quality Index (PSQI), and Behavioral Avoidance/Inhibition (BIS/BAS). During the study, each participant was wearing two devices: a heart rate monitor (Polar H7 heart rate monitor - Polar Electro Inc., Bethpage, NY, USA) to record HR and beat-to-beat interval, and an actigraph (ActiGraph wGT3X-BT - ActiGraph LLC, Pensacola, FL, USA) to capture accelerometer, sleep, and physical activity data. Additionally, the subjects reported all their activities along with their timestamps. Additionally, each one of them filled Positive and Negative Affect Schedule (PANAS) questionnaires at different times of the day (i.e., 10 am, 2 pm, 6 pm, 10 pm, and 9 am the next day), and a Daily Stress Inventory (DSI) questionnaire before they go to bed. Finally, the participants also collected saliva samples twice, one before they sleep and one after they wake up. In summary, for every participant, there is the following information:

- Anthropocentric characteristics (gender, height, weight, age)
- Questionnaire scores (MEQ, STAI-Y, PSQI, PANAS, DSI)
- List of activities (category of activity, time started, time ended)
- Heart Rate (HR) data (IBI, day, time of the day)
- Actigraphy data (accelerometer data, steps, HR, inclinometer data, day, time of the day)
- Sleep data (in bed, out bed, onset, latency efficiency, total minutes in bed, total sleep time (TST), wake after sleep onset (WASO), number of awakenings during the night, average awakening length, movement index, fragmentation index, sleep fragmentation)
- Saliva samples (melatonin, cortisol)

The duration of the study was 24 hours. The participants were required to stay away from drugs for at least a week. Also, everything concerning the data was in accordance with the EU General Data Protection Regulation.

As we show in chapter 2, similar researchers did not have the advantage of using such a feature-rich dataset in their investigations. The various questionnaires helped us correlate different factors between everyday activities, psychological situations, and sleep quality. Also, the saliva samples gave us intelligence about the biological measurements that quantify how good an individual's sleep was. Comparing it with the PSQI questionnaire, we can observe how they perceive their last night's sleep. Moreover, HR data are fundamental during our study on the sleep quality of the participants. In general, there are many experiments one can make to examine the correlation between sleep quality and other factors.

1.3 Objectives

One thing we need to note before presenting the intentions of this project is the definition of *clustering*, which is defined as the technique of grouping instances into subsets based on the similarity of their characteristics [54]. Following this, we discuss the main goals of this thesis.

The first objective of this thesis was to investigate and produce insights into the association between sleep quality and HR data. In order to achieve that, we studied different feature extraction methodologies from time series. Distinctive time series statistics [1], motifs, discords, and snippets [18] are the kinds of features we used to find and match patterns of HR with sleep quality. To discover these patterns, we performed clustering using the extracted features from the HR time series of each user. Moreover, another goal was to detect what daily external factors affect sleep quality. These factors included anthropomorphic measurements, psychological and personality statuses, and activity values. Again, the approach was

similar to the previous; we clustered these features in order to extract information on how external factors are connected to sleep quality. Our final aim was to examine if some of the participants had a false impression of their sleep quality. We proceeded on this matter by combining the outcomes of the previous two clustering tasks. We also used cortisol measurements of each user to strengthen our assumptions.

1.4 Challenges

There is a frequent phenomenon in every thesis that the authors tend to have high expectations during the planning stage. That was also the case in our project. Due to the limited time of 14 weeks in total, we did not have the time to do more extended research between sleep quality, external factors, and pattern recognition. Another reason that aided in the previous matter is that we used a different technique for our proposed problem, and there was not much previous related work (see Chapter 5). Additionally, the theory behind the method we used was highly subject-specific, consuming a considerable amount of time to comprehend it fully before applying it.

Additionally, we were not certain that our proposed methodology would integrate perfectly with the dataset we used. The reason was the limited number of subjects as well as the nature of the data. The MMASH dataset was relatively new, and there were only a few researches applied to it, increasing the level of uncertainty. That means we were not confident about the quality of all the included features except HR, as we show in the next chapter. Chapter 3 presents some data manipulation tasks that had to be done prior to our analysis.

Additionally, the false representation of the population of interest by the data encompassed in the analysis is called *selection bias* or *selection effect*, and almost every real-world dataset is accompanied by it; there was no difference in our occasion. Next, we present the different categories of bias that are related to the MMASH dataset. Firstly, we analyze the sample bias. *Sample bias* occurs when the sample is not randomly selected, meaning that it is more likely to consist of more members with some specific characteristics (e.g., age, gender, preference in food, preference in type of music) than others without these. On our occasion, the sample used was composed of only healthy young males, mainly students from the University of Pisa [55]. That implies that we cannot be assured if our results (see Chapter 4) can represent females of all ages, not even if they are students in the same university. Also, we cannot be sure if our analysis can be generic to the male population, as the majority of our sample comes from people of a specific geographic location. That means that the participants may include different food in their diet, different everyday levels of exercise, in general, different habits. Therefore, we are more confident that our conclusions referred to young males at the University of Pisa. Moreover, the *time interval* occurs when a study is terminated early because the researcher has the desired results. In our case, Rossi et al.'s goal was to provide a one-of-a-kind dataset. However, because of the length of the observations, 24 hours, we can suppose that there is an aspect of time interval bias, and it might have affected our results since people are not behaving the same way each day of the week. Finally, *attrition bias* is caused when participants leave the study before it finishes, which leads to a non-representative sample for the study. Our study was not equipped with attrition bias because the participants were only asked to submit the PANAS questionnaire five times, the DSI questionnaire one time, two saliva samples, and their activities during the day [55]. However, there is always the possibility of the contributors deliberately or not provide untruthful answers or suppress information, which causes the so-called *reporting bias*. That was also something we wanted to investigate, as we mentioned in subsection 1.3. Even though there was a presence of bias in our dataset, it is a natural thing to happen in real-life situations. We presume that our outcomes do not differ significantly from the ground truth. Though, researchers can use the same method on a more generic-population dataset in the future, which could confirm or deny our conclusions.

Last but not least, during the period that this research is taking place, there was a substantial number of COVID-19 cases. The United Kingdom, and England more specifically, was still in lockdown. As expected, that also affected our study as we had to postpone a scheduled meeting. Also, a period like that undoubtedly has influenced our psychological status, and therefore, our performance.

1.5 Summary

To summarize, sleep is extremely important for every living organism. In this project, we aimed to investigate the presence of HR patterns in the different levels of sleep quality. Also, we wanted to explore how daily external factors affect sleep quality, and finally, if there were participants that misjudged or

misinterpreted their quality of sleep. A key advantage for this thesis is the dataset we used and its composition of a wide range of features, from activity statistics to salivary samples. To achieve our goals, several tasks needed to be completed, which are the following:

- Comprehend and apply feature selection for the extracted distinctive time series statistics.
- Understand and deploy different feature extraction methods like motifs, discords, and snippets.
- Find and explain different patterns in the HR data of the participants during their sleep.
- Associate daily external factors with sleep quality and the level of their correlation.
- Combine the results of HR patterns and associated factors to identify users that misjudged their sleep quality.
- Study the theory behind the salivary samples and link it to the participants that misjudged their sleep quality.

The previous list concludes the first chapter of this thesis. Next, chapter 5 provides essential information about the technical aspects of this thesis by reviewing related work. In chapter 3, we introduce the methodology we follow in our problem. Following, in chapter 4, we present and analyze our results. Chapter 5 describes the future work that can be done, and ultimately, in chapter 6, we discuss the conclusions of this project.

Chapter 2

Literature Review

In this chapter, we show related work that has been previously done. We split this section into four parts. The first is about activity recognition and techniques used for time series clustering. Following that, there is an overview of researches that used activity recognition to examine sleep quality, while the third is a synopsis of related work using the MMASH dataset. The latter is a summary of this chapter.

2.1 Feature Extraction from Time Series

We commence by underlining some previous work that was done related to the diverse approaches we use. The undermentioned publications used different features to cluster time series data, which are four in total: (a) distinctive time series statistics; (b) motifs; (c) discords; (d) snippets. We begin with the former.

Abdallah et al. [1] developed an activity recognition framework for mobile phones, which they called STAR. It combined both supervised and unsupervised learning for streaming time series data. They used the first to a pre-obtained dataset to initialize activity clusters, and then with the latter, they were able to generate sub-categories for these clusters. They only saved *key characteristics* for every sub-category like centroid, within sub-cluster standard deviation, and more, using mathematical equations. That was beneficial in two ways: they made the framework run extremely fast, almost real-time, and accurate. The authors also included the personalization factor in that framework by using incremental and active learning. Ultimately, STAR outperformed every other method when they tested it with three different datasets. The exceptional results of this paper motivated us to use distinctive time series statistics to cluster the HR data of the subjects.

Before we proceed with the explanation of the matrix profile, we need to introduce motifs first, which are its inspiration. Patel et al. [48] invented a methodology to discover motifs in time series data. As they quoted, motifs are frequently occurring patterns. A *motif* is the subsequence of a time series sequence with the most matches, and they used Piecewise Aggregate Approximation (PAA), symbolic representations, and Euclidean distance to find it. In our case, though, we used the matrix profile approach to extract motifs. Following, we specify what that is.

Yeh et al. [67] created a novel scalable algorithm to discover motifs and discords for time series, all at once. *Discord* is defined as an unusual subsequence - anomaly in the time series [23]. They designed that algorithm to provide exact results without the need for parameter tuning at exceptional speeds. Regarding the latter, they made it able to be applied on streaming time series as well. It functions by comparing the Euclidean distance of the subsequences between two time series; the second time series can also be the same as the first. The two key elements of the algorithm are the *matrix profile* and the matrix profile index. The first holds the minimum values of the Euclidean distance of a given subsequence and every other subsequence in the time series, whereas the latter has their locations. Finally, when visualizing the matrix profile, values close to zero indicate a motif, while peaks on the plot indicate a discord. In the same year, they presented a faster and more memory-efficient updated algorithm which they tested on a 100-million-length time series dataset, finishing in 12.13 days [71]. Additionally, Zhu et al. [70], building on top of the first matrix profile, introduced a new algorithm that can find motifs even if there are missing data in the time series.

Next, *shapelets* were first introduced by Ye and Keogh [66], and they form the central concept behind snippets. The idea was a small sub-shape of a bigger shape that can provide the maximum information for a class and separate it from others. Zakaria et al. [68] were the first that used shapelets to cluster time series. Their fundamental element was the subsequence distance, which helped them create unsupervised-shapelets (u-shapelets) and a distance map between them and all the time series in the data. After that, they applied the k-means algorithm, getting better Rand index results - a clustering quality metric - compared to other methods in various datasets. Since then, many publications have used u-shapelets, like [43], [15], and [38]. However, because of the lack of source code in Python regarding u-shapelets, we proceeded with snippets.

Imany et al. [28] published a new algorithm based on the matrix profile philosophy in order to discover *snippets* – subsequences of time series. Their initial goal was to find k number of snippets representing the most significant part of the time series. There are two differences between snippets and shapelets. The first is that snippets are, by their nature, unsupervised, meaning we do not need the original labels to extract them. The second distinction between the two was that instead of the Euclidean distance, they used the MPdist [25], which aids in the discovery when the snippet's length or position is not precisely the same as the one compared. Lastly, they presented the overwhelming results of the algorithm by applying it to different noisy datasets, plus two case studies, human behavior, and biology.

Those were the four kinds of features we chose to extract in order to cluster the participants' HR data. We presented our reviews of them along with their benefits. Most importantly, the primary focus of the original researchers during the creation of those approaches was both accuracy and speed. Also, we underline that K-Means was the algorithm used for clustering on one occasion, and Rand index was its quality metric.

2.2 External Factors, Heart Rate, and Sleep Quality

This section is devoted to other researches that linked external factors and HR with sleep quality. For every case, we highlight their methods as well as their results. These will serve us in two ways, guidance and baseline. The first is about gaining intelligence on how other scientists used their data to extract information about sleep quality. Concerning the latter, we use them as a baseline because, as we show in the section 2.3, there has not been a study linked to sleep quality with the dataset we use. Thus, we compare the abstract results between these approaches and ours.

Szöke et al. [61] researched how daily activities affect our sleep. They used FitBit to collect the data from the participants, which they needed to wear at all times. The data included information about burned calories, steps, total distance covered, number of awakenings, time in bed, and more. Then using a linear regression model with the number of awakenings during the night as the dependent variable, they perceived a correlation coefficient of 97%. They concluded that spending much time being very active during the day is related to more awakenings, while on the other hand, being more sedentary is linked to fewer awakenings. Also, the time spent awake has a positive correlation with the number of awakenings, and subsequently, the poor sleep quality.

Sano et al. [58] conducted a 30-day experiment with 66 MIT undergraduate students collecting data using a wrist activity tracker device and a wrist sensor, which they used to find links to academic performance, sleep quality, stress level, and medal health. These devices captured intelligence like three-axis accelerometer data, skin temperature, light exposure levels, and more. They also used an Android app and an MIT website to collect a plethora of other features. Participants also filled out surveys like the Big Five Personality Test and PSQI before and after the research. During the experiment, they reported every morning and evening key information about mood, general health, sleep, activities, social interactions, and more. Concerning our research, they found that PSQI, which lower means better sleep quality, got higher values for high-stress levels and distance traveled and lower for more extensive periods of academic activities.

Bai et al. [6] used directed graphs to predict a participant's sleep quality. They developed an Android app that automatically tracked human activity, location, and environmental information. Moreover, in this app, the 30 individuals submitted their previous' night sleep quality every morning for 30 days. However, the subjects were also required to use another app before sleeping and leave the phone on the bed. The authors found that sleep quality is affected negatively by sitting for a considerable amount of time during the day or exposing to high sound intensities.

Lim et al. [35] developed a smartphone app where the users could provide information based on their physical and social activities, location, and feelings for the past 30 minutes. The 30 participants also self-reported their sleep satisfaction and wore a FitBit tracker for six days. The authors combined

all data, applied Principal Component Analysis (PCA), and finally used Multiple Logistic Regression to classify the five different sleep scales, achieving 0.4574% accuracy.

Burton et al. [11] conducted research regarding sleep quality. They extracted the HR data of 20 individuals, with the majority of them being female. The HR monitoring was done using Polar Heart Rate Monitors (Model S810i, USA; Model RS800, USA), a wrist wearable device manufactured from the same company our participants wore during the experiment, and an HR chest band transmitter. Utilizing linear regression for their analysis, they concluded that low Heart Rate Variability (HRV) is connected to poor sleep quality. HRV is the time between two beat intervals, where lower values translate to higher HR.

In summary, we can see that linear regression is more frequently applied to associate external factors with sleep quality. Additionally, multiple logistic regression was used for sleep classification tasks with the daily activities as the independent variables. Some abstract conclusions propose that increased activity, high-stress levels, distance traveled, sitting for extensive periods, roll-overs, and high-sound intensities negatively affect sleep quality. Finally, a study suggested that high HR is linked with poor sleep quality.

2.3 Previous work using the MMASH dataset

Ultimately, we highlight the literature according to the dataset we use. The dataset was moderately new when this study took place; thus, there were only four publications that cited it during our research. Only the first one used it for a sleep-related study.

Perez-Pozuelo et al. [50] developed an algorithm that uses Heart Rate data from any wearable device with such a sensor to detect sleep. After obtaining the data and preprocessing them to filter out noise, they used different techniques to find and set thresholds about HR values and time. These thresholds aided them in separating a day into big blocks, and after combining it with HR volatility, they classified them as sleep or awake state. Their algorithm had an average sleep time difference of -14.08 minutes in contrast to the ground truth. Furthermore, they applied an angle change algorithm for the accelerometer data for the same task, but the results were not as appealing as the previous method.

Ma et al. [39] used six subjects of the MMASH dataset to evaluate their proposed low-power Heart Rate sensors with the adaptive heartbeat locked loop (AHBLL) technique. Moreover, Morelli et al. [46] used the Heart Rate data from the dataset to accurately predict the standard deviation of inter-beats intervals over 24 hours (SDNN24), in order to estimate the health status of the people that use wearables with HR sensors. Lastly, Rossi et al. [56] also took advantage of the Heart Rate data to measure the effect of missing values on ultra-short HRV data. Additionally, they used them to reproduce findings that claimed to be possible to accurately estimate the root Mean Squared error of successive inter-beat interval differences (rMSSD) and the standard deviation of inter-beat intervals (SDNN) using smaller time windows.

One can quickly recognize that the MMASH dataset was mainly used for health-related studies; however, it was also a tool for one engineering experimentation. Unfortunately, as we mentioned in the section 2.2, none of the studies practiced it for investigating sleep quality. Thence, we do not have a baseline to compare our results against it. Last but not least, we noticed that all of the researchers used the HR data, which implies that they stand as a resilient element of this dataset.

2.4 Summary

To recap, we presented related work with the methodology we used in this project. We introduced the different kinds of features we extracted from the time series, as well as their fundamental differences. Also, we described what algorithms other scientists used for similar projects and what their abstract findings were. Lastly, we displayed the variety of researches that had been done with the same dataset we used.

Chapter 3

Methodology

This chapter describes the approach we follow in this thesis to read, edit, and analyze the data. Hence, this chapter is divided into four sections: the import of the data, the preprocessing steps, the technique to investigate them, and finally, the summary of the chapter. The first step we took during the development was to generate the directories DataFrames and Plots required to save the datasets we created during the data manipulation and plots we use in this thesis, respectively. A (pandas) DataFrame is defined as a table of data [42].

3.1 Importing the Data

We begin by explaining the structure of the original dataset. The provided directory from the original paper consists of a folder called DataPaper along with the published article in .pdf and two .txt files. Inside the DataPaper directory, there is one folder with seven .csv files for every user that participated in the original experiment, 22 in total. However, we eliminated two participants from our analysis. We removed User 1 because he was awake for more than 40 minutes resulting in two sleeping periods and User 11 since he did not have sleep data due to technical problems.

Next, we formed hrDataOfUsers DataFrame with importing each user's HR, day, and time only during his sleeping period. That dataset, after the preprocessing procedure, played the most crucial role in finding HR patterns and linking them to sleep quality. We did not use the accelerometer data because, during our investigation of the actigraph data, we found that a really high percentage of the measurements was 0 during the sleeping period - something expected. Following, the associatedFeatures DataFrame was used for connecting the daily external factors with sleep quality. Lastly, salivaDF was the dataset that composed of the saliva measurements of every user.

3.2 Preprocessing the Data

This section is separated into two subsections, each one corresponding to a specific procedure. The first is handling the missing and unformatted data, whereas the following is the feature scaling.

3.2.1 Handling of Missing and Unformatted Data

Missing and unformatted data are common problems in real-life datasets. Ensuring the quality of data is of the essence prior to analyzing them.

The first thing we discuss is the empty values. Three out of the 20 users in total we used for our analysis had blank cells. The first was user 7, who did not fill his STAY-Y 2 questionnaire, the result of which is a positive integer number. Next, user 18 did not provide his age data, and lastly, we did not have the salivary measurements of user 21 due to the sufficient quality of the sample. We filled the missing data by taking the mean for each column. Regarding the first two cases, we round the outcome to the closest integer. We could also use Machine Learning algorithms for regression to predict the missing values, but we did not want to add bias in our dataset; thus, we proceeded with the safest option.

Subsequently, a couple of users had wrong values in their day column of the actigraph data. Users 8 and 9, instead of having 2 for mentioning the second day of the experiment, as the original paper stated, had -29. We changed these two values manually to the correct ones. Finally, since HR data is a type of time series, the day and time of each observation are extremely important. Hence, we had to manipulate day and time columns; the steps we followed are shown below:

1. We changed values 1 and 2 of the day column to an actual date
2. Merge new day and time columns into one new column
3. Transform the type of the cells from string to DateTime

Additionally, there is one more change we made to the data. PSQI is an index that takes values from 0 up to 21, inclusive. That index separates a poor sleep quality from a good one. Scores over five show that the user did not have a good night's sleep, and vice versa. However, since five is the bound that differentiates the two, we agreed to add another level of granularity in our dataset, the medium sleep quality. Another fact that strengthened our decision is that during our experimentation with the dataset, we found out that 25% of the users had a PSQI equal to five. Thus, since we had three different sleep quality categories, we changed the numerical indexes into sleep quality labels. Table 3.1 shows how we divided the users into three groups.

Table 3.1: PSQI to PSQI Label

PSQI	PSQI Label
>5	High
= 5	Medium
<5	Low

That ends our commentary for the first action. The second, as necessary as the preprocessing step we just analyzed, is feature scaling, and we explain it in the following subsection.

3.2.2 Feature Scaling

As stated in [45], feature scaling prior to clustering returns better quality and more accurate results. Although, before proceeding with this approach and mentioning where we used it, we need to define two terms: standardization and normalization [3], [9].

Standardization is a feature scaling technique that transforms the data in order for their values to have a mean equal to zero and a standard deviation of one. The standardization formula is the following:

$$x' = \frac{x - \mu}{\sigma}$$

, where m is the average and s is the standard deviation across all values, whereas x is the value we want to standardize. Standardization does not have a specific range of values; hence, it does not influence outliers in the dataset. Also, it is advantageous when the data follow the normal distribution.

Next, normalization is the process of transforming the data into a range from 0 to 1. One disadvantage of this procedure is that it affects the outliers. However, it overcomes it as we do not have to make any hypotheses of the data distribution. The equation to calculate the normalization is shown below:

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

, where min characterizes the minimum and max marks the maximum across all values, while x is the value we want to normalize.

Now that we finished with the definition of the two terms, we advance by explaining where we used them and why. We applied the standardization technique to our hrDataOfUsers DataFrame, resulting in the standardizedDataFrame dataset, which is the main contributor to our analysis for finding the HR patterns of sleep quality. We standardized the HR because it follows the normal distribution in healthy adults, like all of our participants, but it also creates outliers in the dataset [34]. Both of these characteristics are associated with standardization. Additionally, normalization was applied to the associatedFeatures DataFrame before we inserted it into the clustering algorithms, which we discuss in section 3.3. As we also mention later, the number of features in the associatedFeatures dataset ranges from

20 to 42. It is easily conceivable that instead of examining each element individually for its distribution, it is more convenient to use the normalization technique. Concerning the last DataFrame we mentioned in section 3.1, salivaDF, we did not use any feature scaling techniques on it because we only used it to plot a graph, as discussed in the next part. That was everything regarding the preprocessing of the data. We continue by explaining the approach we followed for our analysis.

3.3 Analyzing the Data

There are three different objectives we researched, however, this part is segmented into four subsections. The first is about the clustering algorithms we used. The following concerns the discovery of HR patterns and associating them to sleep quality. The third is about researching sleep quality and what daily external factors affect it. The last one regards the combination of the aforementioned two types of research in order to find connections between them.

3.3.1 Clustering Algorithms

This subsection describes the algorithms we used to cluster the data. We present them sequentially as KMeans, Agglomerative Clustering (AC), MyClusterAlgorithm. The latter is an algorithm we created which clusters time series subsequences of equal size into three groups.

KMeans

KMeans [40] is a partitioning clustering algorithm. That means that it groups the data based on the within-cluster variance. While it is one of the simplest and fastest clustering algorithms, it also provides exceptional results. Its procedure is straightforward; given a set of data points drawn from $\Omega = \mathbb{R}^n$, it goes as it follows:

1. Define the k number of clusters for searching.
2. Randomly partition the data points into k sets.
3. Compute the centroid for each subset.
4. For every data point, calculate its Euclidean distance to each centroid and reallocate it to the set corresponding to that centroid.
5. Repeat steps 2 - 4 until no changes take place.

A disadvantage of this algorithm is that it divides the data points into equal cluster sizes [22].

Agglomerative Clustering

On the other hand, AC [47] is a hierarchical algorithm. It builds clusters by minimizing the within-cluster variance, which is similar to KMeans but instead, it uses an agglomerative approach, meaning it creates clusters from individual data points, not the other way around. It stops when all data points are assigned to a cluster.

MyClusterAlgorithm

MyClusterAlgorithm is a function we created that clusters subsequences of the same length in three different groups. It takes two parameters, a dataset with a "User ID" column to extract all users and the equal-sized time series subsequences. The following bullet points demonstrate an abstract way of how this algorithm works.

1. Compute the z-Normalized Euclidean distance of all subsequences between them
2. That creates an (n - 1) x (n - 1) matrix with all these distances where n equals the number of users. Instead of creating an n x n matrix, we remove the first element of rows from columns and the last element of columns from rows. That makes it easier for our task since we do not have duplicate values. All diagonal values beginning from the second row are set to +inf. Table 3.2 illustrates how an imaginary matrix of 5 users would be.

Table 3.2: z-Normalized Euclidean Distances Matrix

	B	C	D	E
A	$\mathbb{R}+$	$\mathbb{R}+$	$\mathbb{R}+$	$\mathbb{R}+$
B	+inf	$\mathbb{R}+$	$\mathbb{R}+$	$\mathbb{R}+$
C	+inf	+inf	$\mathbb{R}+$	$\mathbb{R}+$
D	+inf	+inf	+inf	$\mathbb{R}+$

3. Find the smallest distance in the matrix.
4. If one of the users is already assigned in a cluster, assign the other user to the same cluster.
5. If both users are clustered, do nothing.
6. If none of the users are clustered and there have not already formed three clusters, create a new one. Otherwise, find the next smallest distance that corresponds to a cluster and assign them there.
7. Set the previous value to +inf.
8. Repeat 3 to 7 until all values are +inf.

Having introduced our algorithms, we continue with the introduction of our approach for finding HR patterns in sleep quality. We note that for the implementation of KMeans and AC, we used the sklearn [49] package.

3.3.2 Sleep Quality and Heart Rate Patterns

For this task, we used our standardizedDataFrame, which consists of the standardized HR for every user during his sleep period, the timestamp of each observation, his ID, and PSQI label.

The result of the preprocessing steps of the HR is shown in Figure 3.1. It demonstrates the standardized HR data of every user during their sleep, in descending order by PSQI and colored by PSQI label (see Table 3.1). Also, we need to note that all visualizations in this thesis were created by using two packages, matplotlib [27] and seaborn [64].

It is easily apprehensible that it is almost impossible to find differences between the HR measurements of the three PSQI labels. Some abstract observations are the following:

- The HR of people with good sleep quality is closer to zero for more extensive durations than others.
- The HR of the majority of people with high PSQI is almost constantly over the average of all users.
- There is a positive correlation between the maximum tracked HR and the PSQI label.

Considering the previous were strong assumptions we could make by the eye, we needed to find a way to justify them. In the interest of finding HR patterns based on sleep quality, we thought of utilizing clustering to all users based on their HR. Thus, since our HR data were, in fact, time series, we got inspired to apply feature extraction to them. We used four different approaches for this task:

1. Distinctive Time Series Statistics: statistics of time series, like median, maximum, and length.
2. Motif: repeated pattern in time series.
3. Discord: anomaly in time series.
4. Snippet: unique subsequence of time series.

After we obtained the features, we were able to apply clustering algorithms to them. Following that, we took the predictions of the clustering algorithms and compared them to the reported PSQI labels. Next, we analyzed the ones with the highest similarity to the original PSQI labels.

Furthermore, we used TSLearn's KMeans [62] function. That is an all-around function that extracts features and performs the clustering in the background providing the end-user with the results. What makes it unique is that it can use Dtw Barycenter Averaging (DBA) [51], an averaging method for Dynamic Time Warping (DTW). DTW is a method that computes the similarity between two time series subsequences. DBA is an algorithm that returns the best average subsequence out of a set of subsequences by decreasing the DTW squared distance. The rest of the procedure after obtaining the predictions is the same as the one we mentioned above.

Following, we present how we applied these methods and what parameters were used.

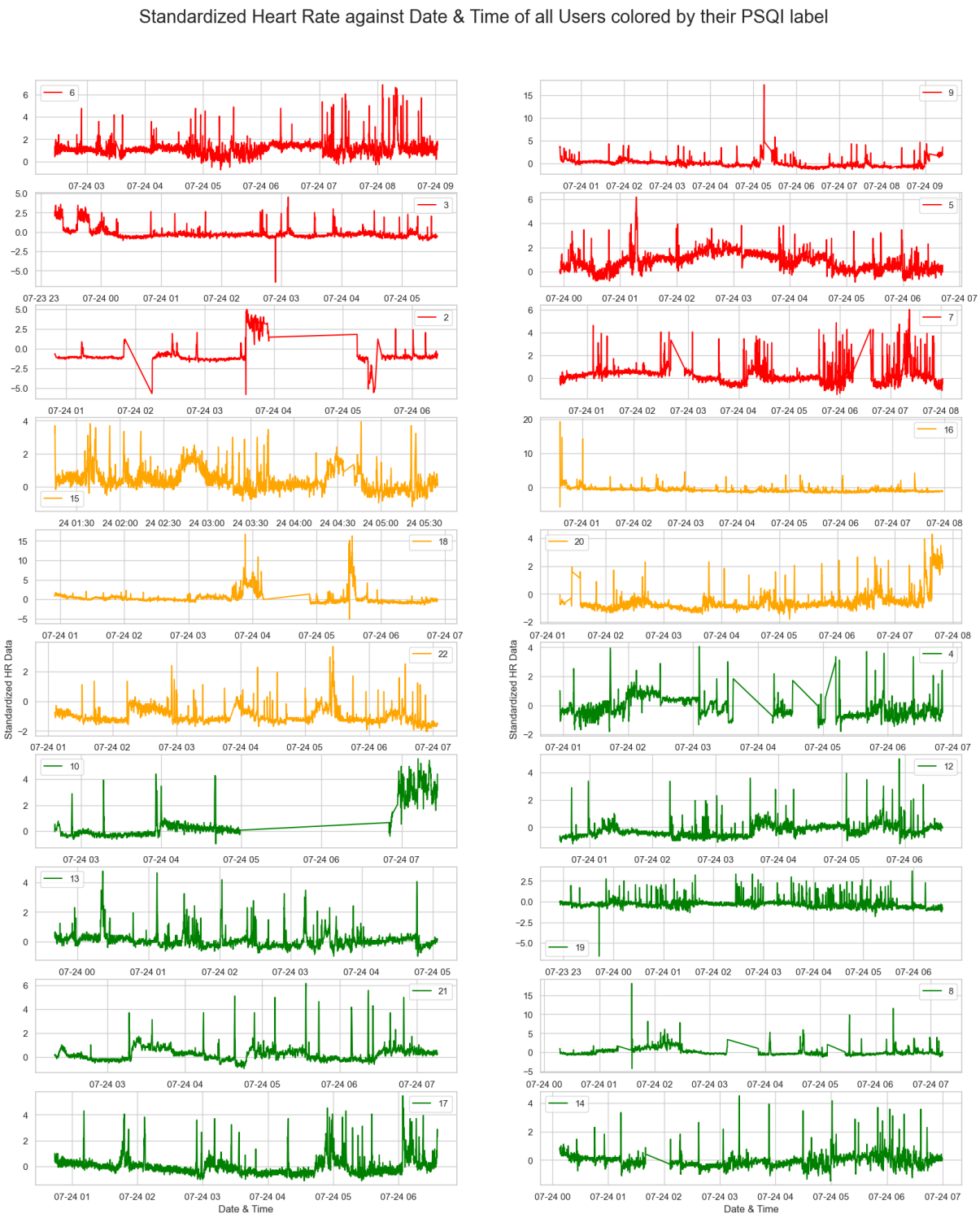

Figure 3.1: Standardized HR against Date & Time for Every User

Distinctive Time Series Statistics

We begin with our first method using the extract_features function from the tsfresh [16] package to extract distinctive statistics of the time series. This function requires three parameters:

1. A pandas DataFrame with the time series.
2. A column_id which corresponds to the column we want our dataset to group by.
3. A column_sort which resembles the column we want to sort by.

We used our standardized DataFrame with the three columns ["User ID", "HR", "DateTime"] as our time series pandas DataFrame, the "User ID" as the column_id, and "DateTime" as the column_sort. The result was a 20 (number of users) x 787 (number of extracted features) DataFrame, which we then passed in the impute function of the same package to deal with the $-$inf, $+$inf, and NaN values. That function replaces these values with the min, max, and median column-wise, respectively. Some extracted features are the mean, the kurtosis, and the sum of reoccurring values of the HR data for every user. We used this DataFrame as an input to two clustering algorithms.

TSLearn's KMeans

For this case, we used from TSLearn's package TimeSeriesKMeans function, which automatically extracts features and performs the clustering in the background. The function takes many parameters, yet we only used three:

1. n_clusters: the number of clusters to form.
2. metric: the barycenter computation method.
3. random_state: seed to ensure reproducibility.

We set the first parameter to three, the second to dtw corresponding to DBA as barycenter computation, and random_state to one. We left the remaining parameters to the default setting. Before fitting our dataset to the model, we had to use the to_time_series_dataset function of TSLearn in order to have the data in a suitable format. Here we used only the HR data from our standardized DataFrame.

Motifs

For the remaining kinds of extracted features, we used the same package, stumpy [33]. That powerful library has been developed based on the Matrix Profile. To begin with, we used the function stump to calculate the matrix profile for every user, one at a time. We used the standardized HR data of each subject as input along with the desired length of the subsequence we wanted to extract. Subsequently, we extracted the motif from the time series in just two lines of code. Finally, after obtaining the motifs for all users, we applied the three clustering algorithms. We repeated this method for four different lengths; 90, 120, 300, and 900 seconds.

Discords

The procedure we followed with discords was very similar to the one with motifs. Likewise, we used the stumpy package, the standardized HR data, the same three algorithms, and finally, the same subsequence lengths.

Snippets

The final feature we used for clustering the sleep quality with the HR is snippets. We utilized the snippets function from the same package by passing three parameters to it:

1. The time series of which we wanted to find the snippets.
2. The desired length of the extracted snippets.
3. The number of snippets we wanted to extract.

The first one was the HR data from our standardized DataFrame for one user at a time; the next was utilized four times with the same length sizes we used for motifs and discords; the latter was set to one, as we wanted to extract only the best snippet for each participant.

Those were the five approaches we used to identify HR patterns and connect them to sleep quality. One last thing to mention before proceeding is that we used a seed wherever possible in order to ensure reproducibility. Furthermore, the code used for this project is available on GitHub [72], along with the instructions to run it. There are also instructions about it in Appendix B. We continue with the plan we used to investigate the daily external factors and how they affect sleep quality.

3.3.3 Sleep Quality and Associated Factors

The approach we followed in order to define the associated factors of sleep quality is similar to the previous. We performed clustering with each user's features, compared them to the actual PSQI labels, and examined the results with the highest clustering accuracy. The examination went thusly:

1. Got the predictions with the highest accuracy
2. Added them as labels for each observation
3. Transformed the problem into a classification task
4. Extracted the coefficients of the features for each class

These coefficients played the role of associating each factor to a specific quality of sleep. Ultimately, we used the original PSQI and transformed the problem into a regression task, getting the correlation between the PSQI value and the features. That way, we were able to evaluate our findings with the dataset's correlation.

We examined three sets of features for this task. In the first, we used all features with the questionnaire raw indexes, a total of the 36 features, consisting of anthropomorphic, psychological, and activity information. Next, we used the original papers' information [55] and transformed the questionnaire indexes into labels. In order for our clustering algorithms to deal with the categorical features, we used the One-Hot Encoding [8] technique. That resulted in an array of 42 features in length. Finally, we removed the personality and psychological dimensions and kept only the body measurements and the activity statistics. The total number of features used in our dataset was 20. For all configurations we used KMeans and AC. Also, before the fitting in the algorithms, we normalized all features column-wise. The list of all features along with their description and units is illustrated in table 3.3 in alphabetic order. We need to note that factors Steps, Standing, Sitting, and Lying, were extracted from the actigraph data, which we calculated manually from the beginning of the day until each participant's sleeping time. Also, we did not use gender as a feature since every participant was identified as male, so it would make no sense to add it.

That was everything regarding associating factors to sleep quality. Next, we explain the purpose we combined the two methods.

3.3.4 Combining The Methods

Since the predictions of the previous clustering algorithms had to do with assigning the participants into groups of different sleep qualities, we combined the results in order to see where these two approaches agree and compared them with the reported PSQI labels. We used the result with the highest similarity between two clustering outcomes. To achieve that, we developed a function that compares every prediction made for sleep quality with each prediction made for associated factors. The purpose was to find users who stated a contrasting sleep quality than the HR data and associated factors suggest. That way, we could tell if a user misjudged his sleep or gave other answers than the truth in the Pittsburgh Sleep Questionnaire. Ultimately, we plotted the saliva data of each user so we could relate our findings with the theory.

3.4 Summary

To summarize, we presented the procedure we followed to import the data, the reasons we did not include all users for our analysis, and finally, why we did not take advantage of the accelerometer data. Next, we defined our preprocessing steps of handling the empty cells, the unformatted data, and the feature

Table 3.3: Description of Associated Factors

Feature	Description	Units
A0	Sleeping	Seconds
A1	Lying down	Seconds
A2	Sitting (e.g., studying, driving)	Seconds
A3	Light movement (e.g., slow/medium walking, work)	Seconds
A4	Medium movement (e.g., fast walking, cycling)	Seconds
A5	Heavy movement (e.g., gym, running)	Seconds
A6	Eating	Seconds
A7	Small screen usage (e.g., smartphone, computer)	Seconds
A8	Large screen usage (e.g., TV, cinema)	Seconds
A9	Caffeinated drink consumption (e.g., coffee, soda)	Seconds
A10	Smoking	Seconds
A11	Alcohol consumption	Seconds
A12	Saliva sample	Seconds
Age	Age of the participant	Years
B_bis	Tendency to perform avoidance behaviors and the sensibility toward negative situations	Score
B_drive	Perseverance and constancy in achieving goals	Score
B_fun	Preference for risky situations, impulsive behaviors, and seeking stimuli that provide immediate and sensory pleasure	Score
B_reward	Predisposition to experience positive effects from reward-related stimuli	Score
DSI	Frequency and magnitude of the stressful events perceived during the day	Score
Evening	If the participant is an evening type. Created by One-Hot Encoding	Binary
G_Average	If the participant was an average anxious type during the test. Created by One-Hot Encoding	Binary
G_High	If the participant was a very anxious type during the test. Created by One-Hot Encoding	Binary
G_Low	If the participant was not an anxious type during the test. Created by One-Hot Encoding	Binary
Height	Height of the participant	cm
Intermediate	If the participant is an intermediate type. Created by One-Hot Encoding	Binary
Lying	How much time a participant was lying before going to sleep	Seconds
MEQ	Morningness–Eveningness questionnaire	Score
Morning	If the participant is a morning type. Created by One-Hot Encoding	Binary
P_neg_10	Level of negative emotions at 10:00	Score
P_neg_14	Level of negative emotions at 14:00	Score
P_neg_18	Level of negative emotions at 18:00	Score
P_neg_22	Level of negative emotions at 22:00	Score
P_pos_10	Level of positive emotions at 10:00	Score
P_pos_14	Level of positive emotions at 14:00	Score
P_pos_18	Level of positive emotions at 18:00	Score
P_pos_22	Level of positive emotions at 22:00	Score
STAI1	Anxiety level during the test	Score
STAI2	Anxiety level in general	Score
Sitting	How much time the participant was sitting before going to sleep	Seconds
Standing	How much time the participant was standing before going to sleep	Seconds
Steps	Number of steps the participant took before going to sleep	Steps
T_Average	If the participant was an average anxious type in general. Created by One-Hot Encoding	Binary
T_High	If the participant was a very anxious type in general. Created by One-Hot Encoding	Binary
T_Low	If the participant was not an anxious type in general. Created by One-Hot Encoding	Binary
Weight	Weight of the participant	kg

scaling. Also, we underlined the logic of why we transformed the PSQI into three different granularities of sleep quality. Lastly, we showed the approach we took to perform the analysis of our data and get to a conclusion.

Chapter 4

Results

This chapter is dedicated to presenting the results of our approach. We demonstrate the outcomes of the various methods we used to cluster the subjects' HR as well as the results from the associated factors' clustering. Also, we demonstrate the aftermath of combining the previous techniques. At the end of this chapter, there is a recap of all sections. Before all that, though, we introduce three terminologies required to interpret the results.

4.1 Statistical Terminology

4.1.1 Rand Index

Rand Index (RI) [49], [69], [52] is a method that computes the percentage of similarity between two arrays or lists of cluster labels by taking into account all pairs of samples and measuring the number of them in the same or different clusters. The actual mathematical equation of RI is:

$$RI = \frac{TP + FN}{TP + TN + FP + FN}$$

A True Positive (TP) is when two labels are in the same cluster, while a True Negative (TN) is when two dissimilar labels are in different clusters. A False Positive (FP) is when two labels are in different classes, and finally, a False Negative (FN) is when two antithetical labels are in two different clusters. Below, we present table 4.1 to clarify how the RI is calculated.

Table 4.1: Rand Index Calculation Table

	Same Class	Different Classes
Same Cluster	TP	TN
Different Clusters	FP	FN

RI ranges from zero to one with a score close to one meaning a perfect clustering result and vice versa. As noted in [41], RI is simply the clustering accuracy.

4.1.2 Multinomial Logistic Regression Coefficients

Contrary to its name, Logistic Regression [60], [10], [19] is used for classification tasks - predicting categorical variables. Thus, the dependent variable is categorical, whereas the independent ones are either continuous or binary. Logistic regression is based on the logistic function (derives from the sigmoid function), and it "uses maximum likelihood estimation to evaluate the probability of categorical membership" [60]. The equation [65] of it is defined as:

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

, where L is the curve's maximum value, k is the logistic growth rate, and x_0 is the x value of the sigmoid's midpoint. Multinomial is an alternation of this function to use more than two dependent

categorical variables. The most significant advantage of this method is that it does not make assumptions on linearity, normality, or homoscedasticity. Figure 4.1 displays the sigmoid function for values from -15 to +15.

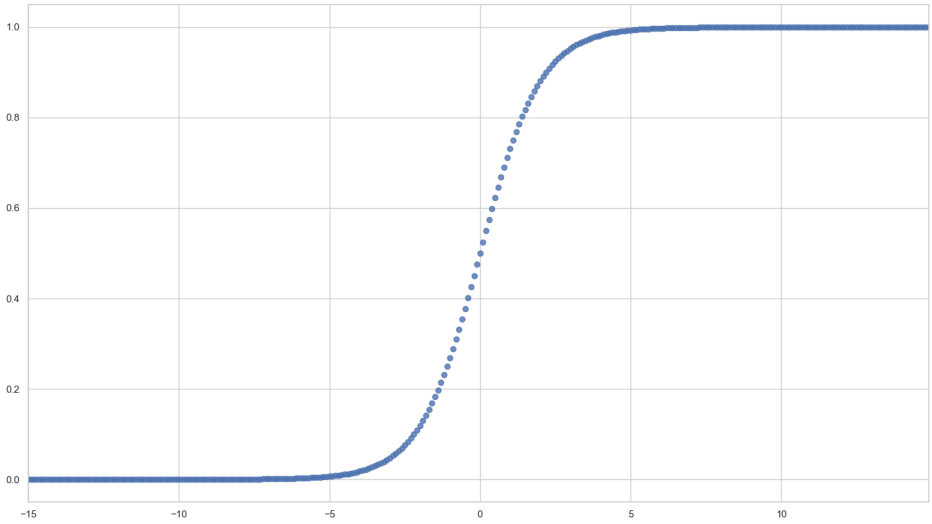

Figure 4.1: An example of sigmoid function

The reason we make use of it is because it helps us estimate coefficients between the dependent and independent variables. These coefficients take values from -inf to +inf, where 0 designates the absence of a relationship between two features. The concept behind our methodology is to use the clustering predictions that returned the best rand score and fit them in a Multinomial Logistic Regressor along with the entire length of features; then, by extracting the coefficients, we will have an indication about the correlation between the dependent variables and the PSQI.

4.1.3 Pearson Correlation

The Pearson Correlation Coefficient (PCC) [37], [19] is a metric that calculates the linear correlation between two variables. It can be calculated from the next equation:

$$\rho_{x,y} = \frac{COV(x,y)}{\sigma_x \sigma_y}$$

, where COV is the covariance and σ is the standard deviation. A couple of its disadvantages is the two assumptions it makes: linearity and homoscedasticity. It takes values from -1 to +1; the first indicates a perfect negative correlation, while the latter implies the opposite. A value close to 0 signifies the absence of correlation. The purpose of using PCC is to validate our findings of the associated factors, which we present later in this chapter.

These three were the metrics needed to comprehend the results of this chapter. We begin by presenting the HR clustering results, followed by the associated factors clustering outcomes. The last in line is an analysis of the combination of these two methods.

4.2 Heart Rate Clustering Results

In this section, we demonstrate the clustering results for sleep quality using the HR data. First, we present the outcome using distinctive time series statistics, followed by the TSLearn's package KMeans implemented method. Succeeding these, we display the results of motifs, discords, and snippets. As noted in the previous chapter, we used the PSQI label of each user as the actual label in order to calculate the RI.

4.2.1 Distinctive Time Series Statistics

The results from utilising this method with KMeans and AC are shown in Figure 4.2. An image that consists of the detailed results of each algorithm is located in Appendix A (Figure A.1).

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Extracted Stats with KMeans	0.4947	1	0	1	1	2	1	2	0	1	1	1	2	2	1	1	0	2	1	1	1
Extracted Stats with AC	0.4947	0	1	0	0	2	0	2	1	0	0	0	2	2	0	0	1	2	0	0	0

Figure 4.2: Rand Score using extracted features with KMeans and AC

One can apprehend that the outcome is not appealing as there is a relatively low rand score (< 0.5). Another thing worth mentioning is that both algorithms predicted the same outcome. Moreover, we observe that Cluster 1 (or 0 for AC algorithm) is the majority prediction in every PSQI label, meaning that most of the subjects had a good sleep quality and did not understand it or the other way around, which does not seem like a logical thing to happen.

For the above reasons, we proceeded by applying feature selection in the previous dataset. Firstly, for this task, we used the sklearn's VarianceThreshold function by setting the threshold parameter to zero. That resulted in removing all columns that had a variance of zero. Subsequently, we used the pandas' corr function to get the correlation between the features. We set a threshold of 0.8 and removed every feature that was highly correlated. After these actions, we had a DataFrame that consisted of 394 columns, removing a total of 393 features - almost half of them. Figure 4.3 shows the clustering results using our updated dataset. Note that there are automated feature selection functions, but they need the labels as another parameter. That would lead to biased results, something that we wanted to avoid.

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Extracted Stats with Feature Selection and KMeans	**0.6105**	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**2**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Extracted Stats with Feature Selection and AC	0.5789	2	0	0	2	1	0	0	2	0	1	2	0	1	0	0	0	0	0	0	0

Figure 4.3: Rand Score using selected extracted features with KMeans and AC

We observe a noticeable increase in the rand score for both algorithms, which was anticipated. Furthermore, even though the two algorithms did not predict the same labels as before, they were highly accurate in predicting the low PSQI users (Cluster 0). Both methods suggest that some of the users with medium or high PSQI share the same attributes as those with low PSQI. However, despite the fact that KMeans had a higher rand score than AC, 0.6105 and 0.5789 respectively, it only clustered one subject as Cluster 2, which is abnormal for this algorithm as it usually returns similar numbers of observations for every cluster. On the other hand, AC predicted more people for each cluster, splitting them between the mid and high PSQI with only one exception, which seems a more robust forecast. Nonetheless, the feature selection method showed a remarkable improvement, suggesting that more thorough research on it could lead to even better results.

Next, we present the TSLearn's KMeans results.

4.2.2 TSLearn's KMeans

Following, we show the results from TSLearn's package TimeSeriesKMeans function, which automatically extracts features and performs the clustering in the background. Below, we can see the predicted labels of the algorithm.

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
TSLearn KMeans with DTW	0.5158	0	1	1	1	1	0	1	2	0	1	1	1	1	1	1	1	1	1	1	1

Figure 4.4: Rand Score using TSLearn KMeans with DTW

As we notice, even though the algorithm returns a rand score of 0.5158, its medium-range value is delusive. That is due to the prediction that 16 out of 20 users to be in the same cluster. There is one possible explanation for this result. The to_time_series_dataset function we used to format the data appended NaN values to the time series in order to match the length of the largest one. One solution could be selecting a fixed length for all time series; an example is the length to be equal to the smallest time series and then use subsequences of the larger ones. However, we did not have enough time to research it since the average time needed for the algorithm to finish was around 12 hours.

4.2.3 Motifs

Figure 4.5 exhibits the outcomes of every algorithm for every length we used.

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Motifs with KMeans and 90s Length	0.6526	1	1	1	0	2	0	0	1	1	2	2	2	2	2	2	2	2	2	1	2
Motifs with Agglomerative Clustering and 90s Length	0.6526	0	0	0	2	1	2	2	0	0	1	1	1	1	1	1	1	1	1	0	1
Motifs with MyClusterAlgorithm and 90s Length	0.6263	1	1	1	2	2	2	2	1	1	0	0	0	0	0	2	2	0	0	1	0
Motifs with KMeans and 120s Length	0.5	2	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
Motifs with Agglomerative Clustering and 120s Length	0.5474	0	2	1	0	2	0	0	1	2	2	2	2	0	2	2	2	2	2	1	1
Motifs with MyClusterAlgorithm and 120s Length	0.5368	2	1	2	1	1	0	1	2	1	1	0	1	0	0	1	1	0	0	2	2
Motifs with KMeans and 300s Length	0.5053	0	1	1	0	1	0	1	1	2	1	1	1	1	1	1	1	1	0	0	0
Motifs with Agglomerative Clustering and 300s Length	0.4632	1	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	1	0	0
Motifs with MyClusterAlgorithm and 300s Length	0.5263	0	1	0	2	0	0	0	1	2	0	2	0	0	0	0	0	0	0	2	1
Motifs with KMeans and 600s Length	0.5789	0	1	2	0	1	2	2	1	2	1	1	1	2	2	2	2	2	1	2	2
Motifs with Agglomerative Clustering and 600s Length	**0.6632**	**1**	**0**	**2**	**1**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**2**	**2**	**2**	**2**	**0**	**2**	**2**
Motifs with MyClusterAlgorithm and 600s Length	0.5684	0	2	1	2	1	0	2	0	2	0	2	0	1	0	1	0	0	2	1	0

Figure 4.5: Rand Score using Motifs

We discern from the previous image a significant improvement in the metric we use. The two best scores of 0.6632 and 0.6526 resulted from AC and KMeans algorithms. We can also identify a better match in the PSQI labels and the clusters. For example, Cluster 2 corresponds to low PSQI with exceptional accuracy in both cases, Cluster 1 to high and 0 to medium. Between the two algorithms, we can see that KMeans created more spherical clusters than AC. To better understand the results, we plot the motifs and how each algorithm clustered them.

By taking a look at figure 4.6, we can see a consistent pattern in Cluster 2. The HR of these users is steady at values close to -0.25 for about nine and a half minutes. That suggests that users 3 and 7 might had a good sleep quality given the motifs of the dataset. On the other hand, we observe that people with medium PSQI had an HR with a bigger variance in contrast with the ones with low. It is also a hint that a couple of subjects that reported a good sleep quality follow patterns with others with worse. Cluster 1 has only two users, which makes it more difficult to discern patterns. Observing it, though, along with the user with high PSQI in Cluster 0, we can assume a disruption at the beginning of the ten-minute period that disturbs sleep. One could state that this might be a false hypothesis as these extracted motifs might start at the end of a good sleep quality period. We believe that this is not the case since the upcoming minutes show an unstable and higher-than the average HR.

Figure 4.7 shows motifs with a length of one and a half minutes. Overall, we can see a really accurate separation of the clusters based on the extracted motifs, with many users following similar patterns. It is hard to discriminate which cluster corresponds to high or medium PSQI users, as in both clusters 0 and 1, the majority consists of subjects with poor sleep quality. However, if we look at the y-axis, we can see that participants at Cluster 0 have an HR higher than the average; combining it with the previous results, we can hypothesize that this cluster corresponds to high PSQI users. As we observed in figure 4.5, subjects of Cluster 2 have an HR close to 0.5, with a peak at the end of the period. Finally, these results could also be the actual labels of the participants, as a self-report questionnaire like the PSQI is vulnerable to people that misperceive their sleep [17].

In general, we can discern how powerful motifs are and their capabilities. Another way of dealing with them is by extracting the K best motifs of each user and clustering using them. That would also erase the possibility that a motif with a high correlation with a specific cluster was not extracted. Next, we discuss discords and their outcomes.

4.2.4 Discords

The image below is the summarized outcome for all executions for discords.

Obviously, discords did not perform as well as motifs; the average rand score with the discords lies at 0.5175, more than 0.05 less than the average of motifs. Surprisingly, we also notice that MyClusterAlgorithm outperformed KMeans and AC on three out of four occasions, while on the other occurrence, they all scored the same. The plot in Figure 4.9 represents the different clusters formatted by the configuration that returned the best rand score, MyClusterAlgorithm with a 120s length size.

An interesting observation is that while we talk about discords, the majority of them have a mean of standardized HR close to zero. Both clusters 0 and 1 share related attributes, with a noisy straight line for two minutes. We can also witness around four outliers in the first two clusters. Participants in Cluster 2, on the other side, have many small ups and downs in their HR during this period. We can recognize that the algorithm groups the subsequences in a logical manner; nevertheless, the separation of the subjects does not seem robust, as there is a mixture of different users with different PSQI labels in the last two clusters. It is evident that compared to the motifs, discords were not really valuable for our

Figure 4.6: Motifs with AC

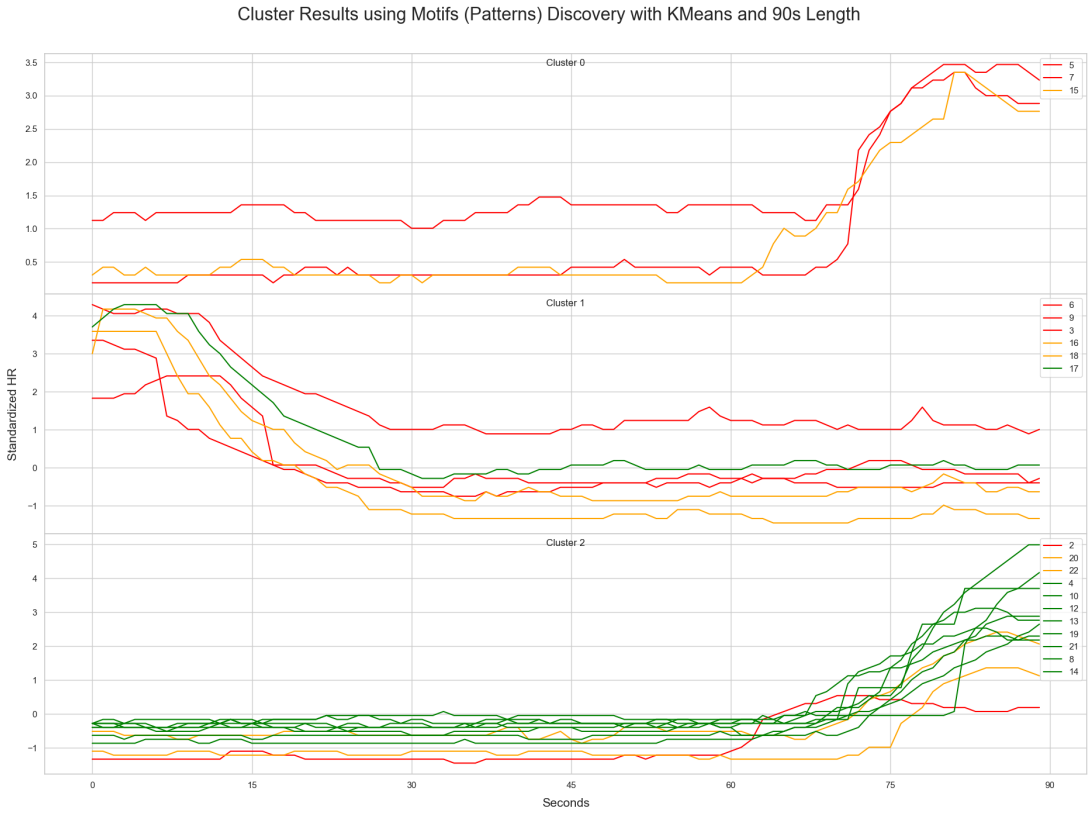

Figure 4.7: Motifs with KMeans

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Discords with KMeans and 90s Length	0.4947	0	0	2	0	2	0	2	0	0	2	2	0	0	2	0	2	2	1	0	0
Discords with Agglomerative Clustering and 90s Length	0.4316	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Discords with MyClusterAlgorithm and 90s Length	0.5684	1	1	1	1	0	0	1	2	2	0	2	2	2	0	1	1	1	0	1	0
Discords with KMeans and 120s Length	0.5158	0	2	2	0	2	0	2	2	1	2	2	2	2	2	2	2	2	0	2	2
Discords with Agglomerative Clustering and 120s Length	0.4632	0	2	2	0	2	2	2	2	1	2	2	2	2	2	2	2	2	0	2	2
Discords with MyClusterAlgorithm and 120s Length	**0.5947**	**1**	**2**	**1**	**1**	**2**	**1**	**1**	**0**	**0**	**2**	**0**	**0**	**2**	**0**	**0**	**1**	**0**	**1**	**0**	**1**
Discords with KMeans and 300s Length	0.5263	0	2	1	0	1	2	2	2	2	1	1	2	2	1	2	1	2	0	1	2
Discords with Agglomerative Clustering and 300s Length	0.5263	0	2	1	0	1	2	2	2	2	1	1	2	2	1	2	1	2	0	1	2
Discords with MyClusterAlgorithm and 300s Length	0.5263	0	2	2	2	0	1	2	0	0	0	0	1	0	0	2	0	0	1	0	2
Discords with KMeans and 600s Length	0.5158	2	2	0	2	0	2	2	0	0	0	0	0	0	2	2	0	0	1	0	2
Discords with Agglomerative Clustering and 600s Length	0.5158	1	1	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	2	0	1
Discords with MyClusterAlgorithm and 600s Length	0.5316	2	1	0	1	2	2	2	0	0	1	2	1	0	1	2	2	1	2	2	2

Figure 4.8: Rand Score using Discords

occasion. Extracting the K best discords might probably return better results, but this will be a task for future work.

Last but not least, we demonstrate the results by extracting snippets and then clustering them.

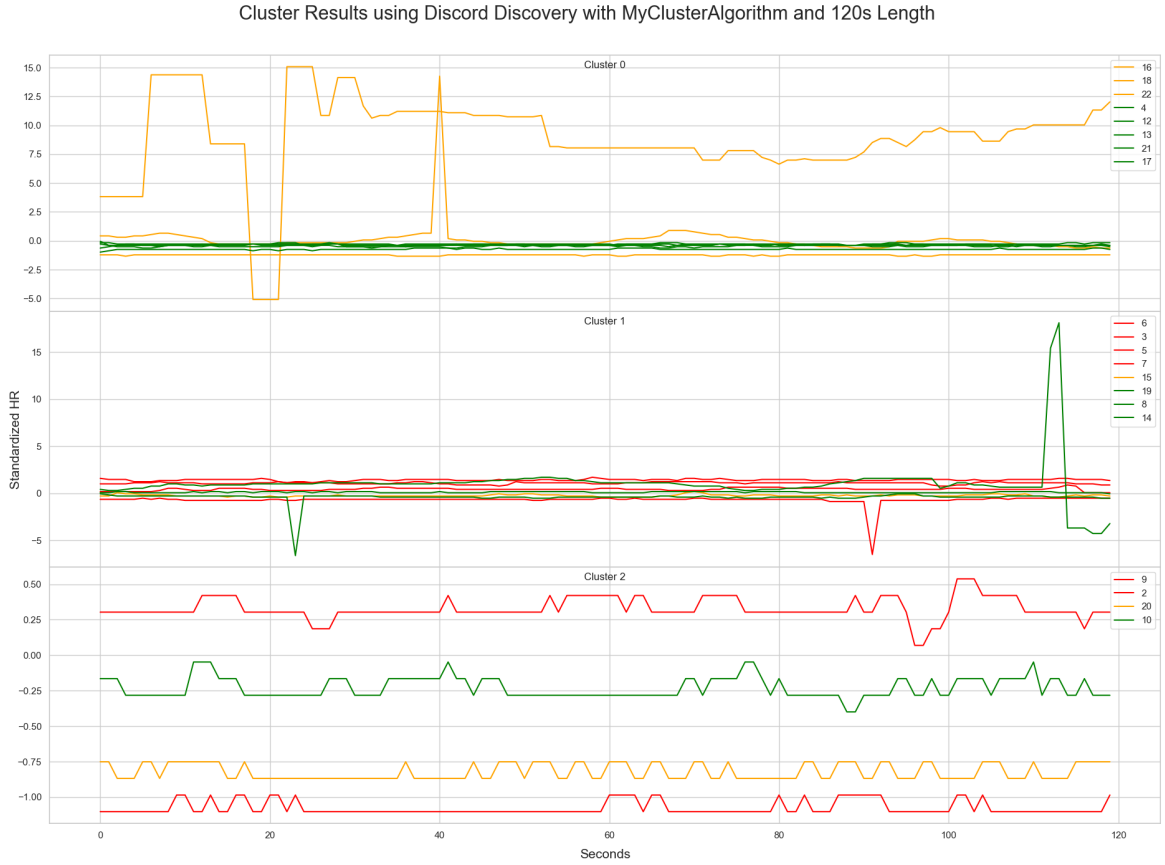

Figure 4.9: Discords with MyClusterAlgorithm

4.2.5 Snippets

The detailed outcome for each configuration for Snippets is shown in Figure 4.10.

The results reveal an improvement over the discords and an average rand score of 0.5521, close to the average of the motifs (0.5684). Remarkably, MyClusterAlgorithm returned the best outcome and outperformed the other two algorithms on three out of four occasions. We distinguish at the last row a good division of the subjects with low and medium PSQI, while high PSQI participants seem to have been separated in a mixture of classes. To better interpret this result, we display the plot of snippets by clusters with MyClusterAlgorithm and length of 10 minutes in Figure 4.11.

As we witnessed with motifs, people with good sleep quality (see Cluster 1) tend to have a constant standardized HR close to zero for more extended periods of time than the other participants. There is also evidence of a slight increase over time. Cluster 0, which resembles subjects with high PSQI, shows

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Snippets with KMeans and 90s Length	0.5368	0	1	1	1	2	1	1	2	1	2	2	2	0	1	1	2	2	0	1	2
Snippets with Agglomerative Clustering and 90s Length	0.5158	2	0	0	0	1	0	0	1	0	1	1	1	0	0	0	1	1	0	0	1
Snippets with MyClusterAlgorithm and 90s Length	0.5526	0	1	1	0	1	0	0	2	0	1	0	0	0	0	0	2	0	2	2	0
Snippets with KMeans and 120s Length	0.5947	0	1	1	0	1	2	2	2	0	1	2	2	0	2	0	2	2	2	0	2
Snippets with Agglomerative Clustering and 120s Length	0.6	0	1	1	0	1	2	0	2	0	1	2	2	0	2	0	2	2	2	0	2
Snippets with MyClusterAlgorithm and 120s Length	0.5526	2	1	1	2	1	2	2	0	2	1	2	2	2	2	2	0	2	0	0	2
Snippets with KMeans and 300s Length	0.5474	2	2	2	1	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
Snippets with Agglomerative Clustering and 300s Length	0.4842	0	0	0	1	2	0	0	0	0	2	2	0	0	0	0	2	0	0	2	0
Snippets with MyClusterAlgorithm and 300s Length	0.6105	2	1	0	2	1	2	0	1	0	1	0	2	2	2	2	1	0	2	1	2
Snippets with KMeans and 600s Length	0.5158	1	0	0	1	2	0	0	0	1	0	0	1	1	0	1	0	0	1	1	0
Snippets with Agglomerative Clustering and 600s Length	0.4947	1	0	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
Snippets with MyClusterAlgorithm and 600s Length	**0.6211**	**1**	**0**	**2**	**1**	**1**	**0**	**0**	**2**	**2**	**2**	**2**	**1**	**1**	**1**	**1**	**2**	**0**	**1**	**2**	**1**

Figure 4.10: Rand score using Snippets

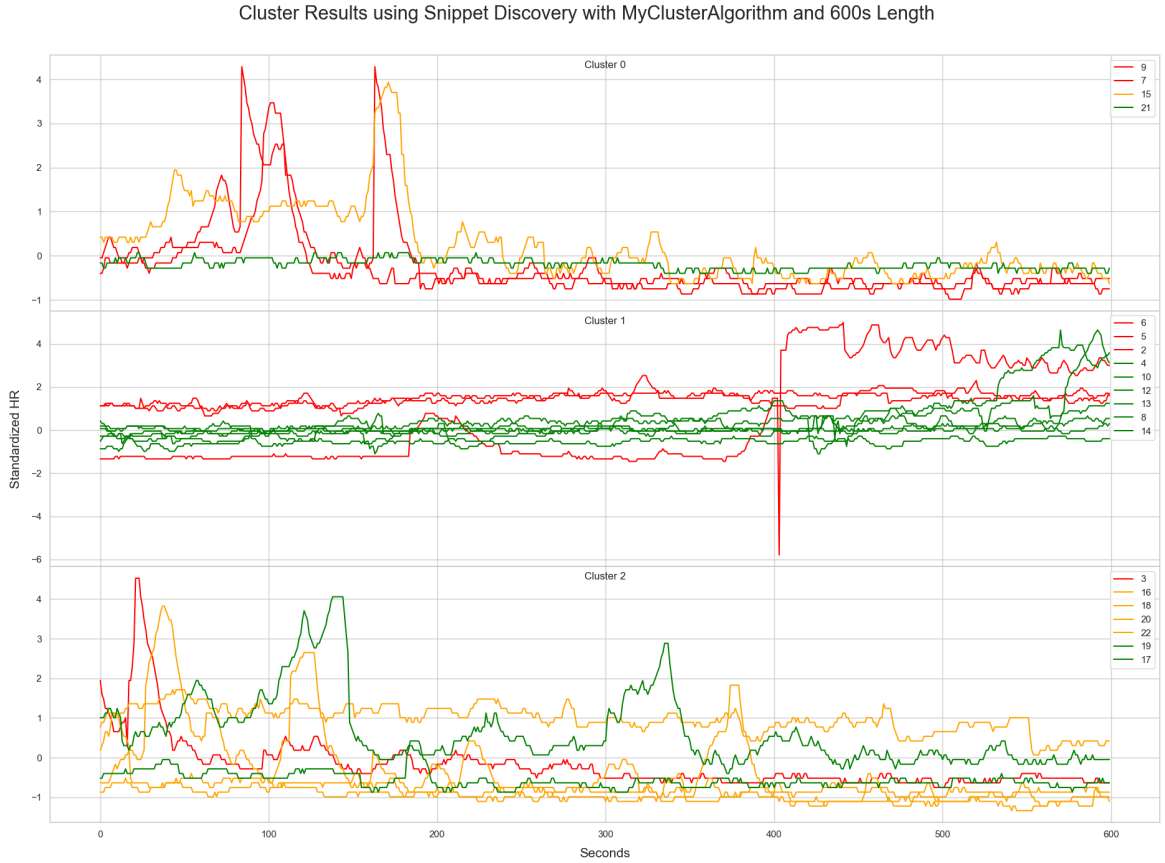

Figure 4.11: Snippets with MyClusterAlgorithm

a pattern of a couple of big spikes in the HR of the participants and a normalization close to zero after that. Surprisingly, people that reported a medium sleep quality seem to have a more unstable HR than people with high PSQI. Moreover, the average standardized HR of theirs is farther away than 0.5 and -0.5, which is not the case in the other two clusters. We can also discern a decrease in the standardized HR of people with medium PSQI over time. Additionally, we can see a user that reported good night sleep but shares more similarities with people with medium PSQI. The same applies to two participants in Cluster 1, which show a constant HR in a ten-minute period.

Generally, snippets provided us with valuable information about the association between HR and sleep quality. Because of the nature of a snippet, meaning a distinctive subsequence of a time series that makes it unique, further research with a longer length as a parameter can be done. Also, clustering the K best snippets could produce better results.

That was the last method we used to cluster sleep quality with HR. Throughout this subsection, we witnessed that the last three kinds of extracted features provided us with essential knowledge that relates HR to sleep quality. We also witnessed the power of motifs and how they scored higher than every other method and even the distinctive statistics of time series. Next, we continue with the results of the clustering of associated factors concerning sleep quality.

4.3 Associated Factors Clustering Results

This section emphasizes the clustering outcomes using only the associated factors of every user. We split this segment into three parts of clustering results; (a) associated factors using questionnaire index results; (b) associated factors by transforming questionnaire raw results to labels; (c) associated factors using only anthropomorphic and activity stats. As mentioned in Subsection 3.3.3, we used KMeans and AC as our clustering algorithms.

4.3.1 Questionnaire Raw Indexes

Figure 4.12 displays the clustering results. Also, the summarized clustering results of all configurations can be found in Appendix A (Figure A.2).

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Ass. Factors Using Questionnaire Results with KMeans	0.4526	0	0	0	0	0	1	0	0	1	0	0	0	2	1	0	1	0	0	0	1
Ass. Factors Using Questionnaire Results with AC	**0.5789**	**2**	**2**	**2**	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**1**	**2**	**1**	**0**	**0**	**0**	**2**	**2**	**2**	**0**

Figure 4.12: Rand score using Questionnaire Raw Results

We observe two totally different outcomes from the two algorithms, with the average rand score of the two equal to 0.5158. KMeans, in contrast to its nature, used only one participant for an entire cluster. Contrarily, the AC algorithm separated the different participants in a more appropriate way; Cluster 2 - high PSQI, Cluster 1 - medium PSQI, Cluster 0 - low PSQI. However, there is a high presence of people with poor sleep quality in people with low PSQI. The following graph shows the 15 most correlated variables with each PSQI label, produced by the Multinomial Logistic Regression's coefficients, using AC's predictions as class labels. The features are gradient colored from red to green, corresponding from the lowest to the highest value.

	A0	P_neg_10	P_neg_14	STAI1	A8	P_neg_18	P_pos_22	Sitting	Lying	P_pos_10	A11	A2	P_neg_22	A1	P_pos_14
Cluster0	-0.73914	-0.51789	-0.49471	-0.42294	0.410165	-0.40484	0.398887	0.380411	-0.37574	0.349458	-0.32128	0.307196	-0.29334	-0.28234	0.248134

	P_neg_10	STAI1	P_neg_14	P_neg_18	Height	Weight	B_fun	B_reward	A8	A1	Standing	P_neg_22	A12	A10	A9
Cluster1	0.685947	0.648558	0.538844	0.493425	0.373568	0.351408	0.292237	0.290072	-0.28283	-0.242	0.238465	0.223931	-0.22072	0.20843	0.200384

	A0	P_pos_22	A1	Sitting	A9	Steps	P_pos_14	Standing	P_pos_10	A5	P_pos_18	A7	B_fun	STAI1	Lying
Cluster2	0.774736	-0.55374	0.524341	-0.462	-0.35618	-0.34983	-0.33818	-0.32688	-0.30974	-0.28979	-0.27242	0.244853	-0.24195	-0.22562	0.224956

Figure 4.13: Multinomial Logistic Regression's coefficients of ass. factors using questionnaire results with AC

Interestingly, while a high positive correlation exists between A0 (the full explanation of each feature is available at table 3.3) and poor sleep quality, it is highly negatively correlated with low sleep quality. That suggests people with more caffeinated consumption throughout the day had a worse sleep quality than others who did not. Moreover, most of the highly correlated factors with Cluster 1 are either psychological or anthropomorphic. By looking at the four related factors of Cluster 1, we can tell that

people with medium PSQI had high negative emotions during their day. On the same wavelength, people with fewer positive feelings before sleep had a high PSQI label - a very legitimate consequence. On the other hand, fewer negative feelings during the day provided a better night's sleep.

In the following subsection, we present the results of the converted psychological indexes into labels.

4.3.2 Questionnaire Labels Results

The image below shows the results we obtained using One-Hot Encoding to questionnaire raw indexes.

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Ass. Factors Using Questionnaire Labels with KMeans	0.5053	1	1	2	0	0	1	0	0	1	1	1	2	1	0	0	0	1	1	1	2
Ass. Factors Using Questionnaire Labels with AC	**0.5316**	**0**	**0**	**1**	**0**	**2**	**0**	**2**	**2**	**0**	**0**	**0**	**1**	**0**	**2**	**2**	**2**	**0**	**1**	**0**	**1**

Figure 4.14: Rand score using Questionnaire Labels Results

We perceive a slight increase in the KMeans rand score, yet both algorithms did not match or pass the previous rand score. However, the mean rand score of both algorithms is 0.5185, a value close to the previous result. Likewise, the best-performed algorithm here is AC, which distributed the users with low PSQI labels equally into three clusters. It seems like Cluster 0 corresponds to users with high PSQI labels, Cluster 1 to low, and Cluster 2 to medium. Figure 4.15 demonstrates the 15 most effective coefficients using the Multinomial Logistic Regression.

	T_Low	T_Average	Morning	A1	ntermediat	A0	P_neg_10	Sitting	B_bis	A11	P_pos_10	P_neg_18	Steps	Height	A6
Cluster0	-0.79803	0.764967	-0.5411	-0.44043	0.424801	-0.31887	0.313735	0.302258	-0.25059	0.236494	-0.20788	0.174498	-0.17384	0.162853	0.161936

	Morning	ntermediat	P_neg_18	B_reward	Age	A0	P_pos_10	P_neg_22	B_fun	A7	A2	A8	Sitting	A12	A6
Cluster1	0.924353	-0.85184	-0.26788	-0.26642	0.251689	-0.24215	0.216471	-0.17687	-0.14352	0.140206	0.13947	0.134068	-0.13299	-0.12231	-0.11725

	T_Average	T_Low	A0	ntermediat	A1	Morning	B_bis	A12	P_neg_10	A7	Standing	B_reward	G_Average	G_Low	Sitting
Cluster2	-0.7269	0.686051	0.561024	0.42704	0.396412	-0.38326	0.311581	0.242127	-0.23589	-0.2315	0.217009	0.197378	0.194915	-0.19488	-0.16927

Figure 4.15: Multinomial Logistic Regression's coefficients of ass. factors using questionnaire labels with AC

Before analyzing the outcome, we need to mention that because of the One-Hot Encoding, it is normal to see opposite columns next to each other. For example, T_Low that corresponds to the low anxiety during the time of the test is contrary to T_average.

In contrast with the previous results, the ones above (Figure 4.15) suggest that the individuals' personalities and psychology play the most critical roles in their sleep quality. Participants with moderate anxiety during the test have a worse sleep than others with low. Similarly, morning type persons had a better night's sleep than the people whose "circadian rhythm produced peak alertness between the morning and the evening" [5]. Additionally, in terms of activities, people that sleep and lie down more during the day show a worse sleep quality than participants who do not. That observation was similar to the one we did in the previous subsection. Overall, this configuration did not return remarkable results; nevertheless, it provided insight into the different factors associated with sleep quality.

In the previous two outcomes (Figure 4.13 and 4.15), we detected the importance of the diverse activities during the day. Thus, we proceed with our last structure of features, which consists of only anthropomorphic and activity statistics.

4.3.3 Activity Statistics Results

The segmentation of the users into clusters of the two algorithms, using only anthropomorphic measurements and activity statistics, is shown below in Figure 4.16.

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Ass. Factors Using Only Activity Stats with KMeans	**0.5895**	**0**	**0**	**2**	**1**	**1**	**0**	**2**	**2**	**1**	**2**	**1**	**0**	**1**	**0**	**0**	**0**	**2**	**2**	**2**	**0**
Ass. Factors Using Only Activity Stats with AC	0.5316	0	0	1	1	1	2	1	0	0	0	1	0	1	2	2	2	1	1	0	0

Figure 4.16: Rand score using Activity Statistics Results

As we can notice, the average rand score of the two algorithms, 0.5606, is at least 8% improved compared to the previous outcomes (Figures 4.13 and 4.15). Opposing those results, we observe that KMeans outperformed AC and also returned the best rand score across all feature configurations. The value of 0.5895 is nowhere near the best rand score we achieved with the feature extraction of the

standardized HR data; however, these features compose the associated factors, and they have a non-direct relation to the actual sleep quality. The KMeans result from the previous image suggests that Cluster 0 corresponds to good sleep quality users, Cluster 1 to poor sleep quality, whereas Cluster 2 resembles medium sleep quality. The following chart (Figure 4.17) represents the 15 most significant coefficients produced by the Multinomial Logistic Regression using the labels KMeans predicted and the 20 dimensions we reported earlier.

	A8	A9	A0	A6	Weight	A11	A1	A7	Sitting	Height	Age	A5	A4	A3	A2
Cluster0	0.692954	-0.64323	0.62062	0.583615	-0.45014	-0.42338	0.398023	0.367956	-0.33584	-0.32328	-0.31221	-0.26312	-0.24739	0.157312	-0.15094

	A9	A7	A5	A6	Weight	Age	A2	Height	Sitting	Steps	A0	Standing	A10	A4	A1
Cluster1	1.079853	-0.6704	0.63251	-0.39655	-0.34609	0.311269	0.308558	-0.29899	0.275842	0.259933	-0.25427	0.199066	-0.17502	-0.16655	-0.12394

	Weight	A8	Height	A9	A4	A5	A0	A11	A7	A10	A1	A6	A2	A3	Steps
Cluster2	0.796234	-0.66319	0.62227	-0.43663	0.413942	-0.36939	-0.36635	0.348645	0.302443	0.299113	-0.27408	-0.18707	-0.15762	-0.1536	-0.13652

Figure 4.17: Multinomial Logistic Regression's coefficients of ass. factors using activity statistics with KMeans

Concerning Cluster 0, we discern a medium positive correlation between time spent with large and small screen usage, eating, and contrarily to our previous findings, sleeping and lying down. The negatives are caffeinated and alcohol consumption, sitting, and all the body measurements. Most of these results genuinely make sense. For example, when people eat more, they feel more tired and want to sleep, or when people drink more coffee, they want to stay up late, and consequently, sleep fewer hours.

Regarding Cluster 1, we observe a strong positive correlation with caffeinated consumption and a medium positive correlation with heavy movement, age, and sitting. On the other hand, there is a negative correlation between small screen usage, eating, and weight. These results are also legitimate. People consume caffeine to obtain more energy and stay awake for extended periods, while more prolonged mobile phone usage worsens sleep quality [57]. Another reason that makes this outcome consistent is the values of heavy movement, eating, weight, and steps. It is common knowledge that when someone exercises more and eats less, he loses weight.

Sequentially, about Cluster 2, we perceive a positive correlation with height, medium movement, alcohol consumption, small screen usage, and weight, with the last one having a strong and the rest medium correlation. On the opposite, there is a medium negative correlation with extensive screen usage, caffeine consumption, heavy movement, and sleeping. It is difficult to explain what medium sleep quality is as it is something we invented more or less. However, we can see that this cluster shares patterns with the other groups of people with poor and good sleep quality. Alcohol and caffeine consumption contract with Clusters 0 and 1, respectively, while small screen usage and sleeping are two agreements with these Clusters. That was not the case with Clusters 0 and 1, as the only similarities they shared were the negative correlations with weight, height, and medium movement out of the 15 most significant dimensions for each one.

To relate these results to the ground truth, we calculate the PCC of the features we used for our last experiment. Since the PCC cannot deal with categorical variables and we also want to examine our results compared to the actual ones, we create a new column with the actual PSQI and drop the PSQI label. That way, we transform our problem as it is a regression. After executing the code to calculate the PCC for the actual PSQI, acquired the following 15 most correlated features (Figure 4.18).

	Lying	A1	A9	A3	A2	A6	Age	Height	Steps	A10	A0	A11	Sitting	Weight	A8
PSQI	-0.43647	-0.37817	-0.34033	0.296264	-0.27336	0.200753	-0.19121	0.171102	0.166696	-0.14361	-0.0987	-0.08843	0.088061	0.076883	0.068305

Figure 4.18: PCC of ass. factors using activity statistics with the actual PSQI

In order to analyze the results correctly, we compare them with Cluster 0 from Figure 4.14, which characterizes our lower bound.

Firstly, we see Lying as the most negatively correlated feature, which was not present in the 18 unique we computed with the Multinomial Logistic Regression. Unfortunately, we see A1 being opposite to the analog of Cluster 0's. However, the following five features show a similar correlation with Cluster 0 in terms of the sign. For example, as expected, A9 (caffeinated drink consumption) is negatively related to PSQI. By inspecting Cluster 1, we can comprehend that Steps are also positively correlated with PSQI. Height, though, is the second feature that does not follow our pattern. The rest of the features have little impact on the PSQI ($\rho < 0.15$). Overall, we can notice that the results we discussed earlier do not differ significantly from the ground truth.

Ultimately, even though we did not achieve a high rand score, we obtained valuable, and most of all consistent, insight about the associated factors and how they relate to sleep quality. Following next, we perform cross-validation between our clustering results from sleep quality and associated factors.

4.4 Heart Rate and Associated Factors Combination

This subsection is devoted to combining the results of our preceding two clustering results. Rand score is used again as the metric we use to define the similarities between two clustering outcomes. The following chart (Figure 4.19) shows the top 20 combinations out of 245, sorted by their rand score in descending order. The cells of each of the first two columns are deliberately colored based on the method used so that we can process faster the frequency of each method.

Sleep Quality Method	Associated Factors Method	Rand Score
Motifs with MyClusterAlgorithm and 90s Length	Ass. Factors Using Only Activity Stats with AC	0.6316
Motifs with MyClusterAlgorithm and 600s Length	Ass. Factors Using Only Activity Stats with KMeans	0.6211
Snippets with MyClusterAlgorithm and 300s Length	Ass. Factors Using Only Activity Stats with KMeans	0.6211
Motifs with MyClusterAlgorithm and 90s Length	Ass. Factors Using Questionnaire Labels with KMeans	0.6158
Discords with KMeans and 90s Length	Ass. Factors Using Only Activity Stats with AC	0.5947
Discords with MyClusterAlgorithm and 300s Length	Ass. Factors Using Questionnaire Labels with AC	0.5947
Motifs with MyClusterAlgorithm and 120s Length	Ass. Factors Using Only Activity Stats with AC	0.5947
Motifs with MyClusterAlgorithm and 90s Length	Ass. Factors Using Questionnaire Results with AC	0.5947
Motifs with MyClusterAlgorithm and 600s Length	Ass. Factors Using Only Activity Stats with AC	0.5947
Discords with Agglomerative Clustering and 600s Length	Ass. Factors Using Only Activity Stats with KMeans	0.5895
Discords with KMeans and 600s Length	Ass. Factors Using Only Activity Stats with KMeans	0.5895
Discords with MyClusterAlgorithm and 300s Length	Ass. Factors Using Questionnaire Labels with KMeans	0.5895
Snippets with MyClusterAlgorithm and 600s Length	Ass. Factors Using Only Activity Stats with KMeans	0.5895
Motifs with KMeans and 600s Length	Ass. Factors Using Questionnaire Results with AC	0.5895
Extracted Stats with Feature Selection and AC	Ass. Factors Using Only Activity Stats with KMeans	0.5789
Motifs with MyClusterAlgorithm and 90s Length	Ass. Factors Using Questionnaire Labels with AC	0.5789
Motifs with MyClusterAlgorithm and 120s Length	Ass. Factors Using Questionnaire Labels with KMeans	0.5789
Snippets with MyClusterAlgorithm and 300s Length	Ass. Factors Using Only Activity Stats with AC	0.5737
Motifs with KMeans and 90s Length	Ass. Factors Using Only Activity Stats with AC	0.5737
Motifs with Agglomerative Clustering and 90s Length	Ass. Factors Using Only Activity Stats with AC	0.5737

Figure 4.19: Rand Score of combining sleep quality and associated factors clustering results

One can apprehend that motifs is the method that has the most common results with the associated factors. Surprisingly, discords have a greater presence than snippets in this plot, something that was not expected based on the results we presented earlier. The extracted statistics and TSLearn's KMeans provided the most unsatisfactory results. Concerning associated factors methods, the configuration with the anthropomorphic measures and activity statistics delivered the best results. Astonishingly, the algorithm that we developed was the one that produced the best outcome for sleep quality, whereas AC was the one for the associated factors. Both of these configurations did not provide the best rand score when compared to the PSQI labels. In terms of the rand score, we witness a substantial value of 0.6316.

We continue by presenting in Figure 4.20 the two clustering predictions side-by-side in order to analyze them for similarities. We also colored the predictions of the algorithms based on their frequency in each different PSQI label, starting from the one with the more significant presence in a PSQI label. Green corresponds to low PSQI values, yellow to medium and red to high. We grouped them in that manner because these two algorithms returned a high rand score with the original PSQI labels, and that is the only logical way that represents that result.

Method	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Motifs with MyClusterAlgorithm and 90s Length	1	1	1	2	2	2	2	1	1	0	0	0	0	0	2	2	0	0	1	0
Ass. Factors Using Only Activity Stats with AC	0	0	1	1	1	2	1	0	0	0	1	0	1	2	2	2	1	1	0	0

Figure 4.20: Best combined Sleep Quality and Associated Factors methods and their predictions

Our task in this subsection is not to compare the results to the original PSQI labels; we want to find the connections between the two. We can see that the two algorithms agree that users 6, 9, 15, and 17 do not belong to the group they reported in their Pittsburgh Sleep Quality Questionnaire.

The only way to examine these allegations is by plotting the features of these users alongside the others. Thus, Figure 4.21 shows the 90-second length motifs and how they were clustered by MyClusterAlgorithm, whereas Figure 4.22 is a parallel coordinates plot. The latter is an advantageous technique to plot a dataset when it has various dimensions. For a better comprehension of the outcomes, we used different colors for the previously mentioned users.

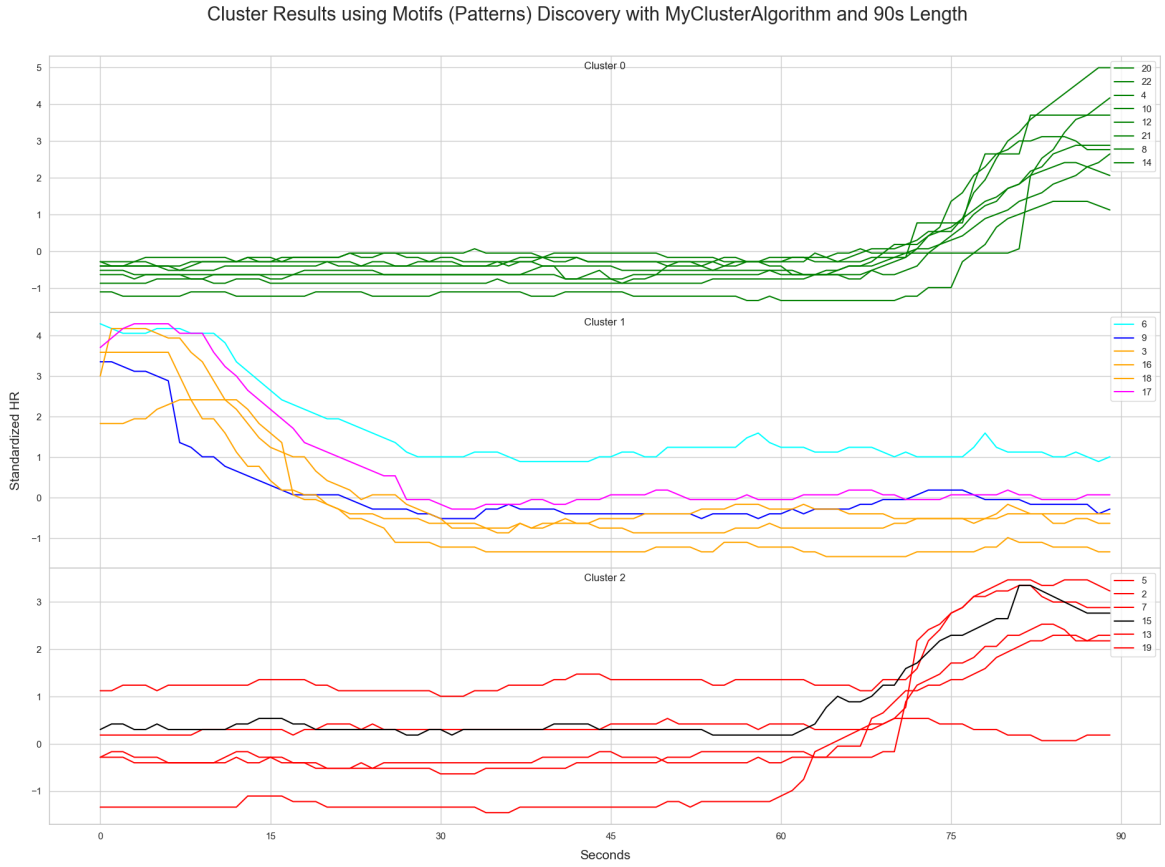

Figure 4.21: Motifs of 90s length clustered with MyClusterAlgorithm

About motifs, we observe an expected result. Our algorithm is based on the distance between the subsequences; hence, the projected motifs of the users most definitely will have a better match with the predicted labels. We witnessed before that motif was the feature that provided the highest rand score. However, since we only use the best motif in the time series, it is like setting a unique identity for every user. We observe distinctive patterns in each cluster. For example, Cluster 1 has a local maximum in the first few seconds of the one-and-a-half-minute period, with a stabilizing of the HR after 30 seconds. Cluster 0 shows a constant HR until the 73rd second, where there is a monotonous increase, reaching a local maximum at the end of the motif. Finally, Cluster 2 shows an HR increase at the 67th second, achieving a local maximum close to the 80th second and then decreasing until the end of the motif. This information is precious and directly connected to sleep quality; however, more research needs to be done on this topic by extracting more motifs and then comparing them with the PSQI labels and our findings.

Following this, we analyze the parallel coordinates chart for each user. Before that, though, we present some generic conclusions we made from the plot. Clusters 0 and 2 take low anthropomorphic values, while Cluster 1, even though it shows a big variance in these values, most of them are more significant. Cluster 2 keeps taking low values for the following four features, Cluster 1 continues with the high values, and Cluster 0 shows a slight gain. Concerning the activities, Cluster 0 records increased A1, A6 to A9, A11, A12, average values in A3, and finally, low values in A2 to A5, and A10. Cluster 1 displays high values in A2, typical A3, A6, A7, A9, and low values to the rest. Finally, Cluster 2 exhibits increased A0 and A6, average A1, A3, A8, with the remaining being close to zero.

With the previous statements, we begin with user 6, who reported poor sleep quality. We see that the first seven dimensions are more like a good sleep quality pattern; from A0 to A8, he follows a similar style with a medium sleep quality participant, while the remaining features are closer to a low PSQI

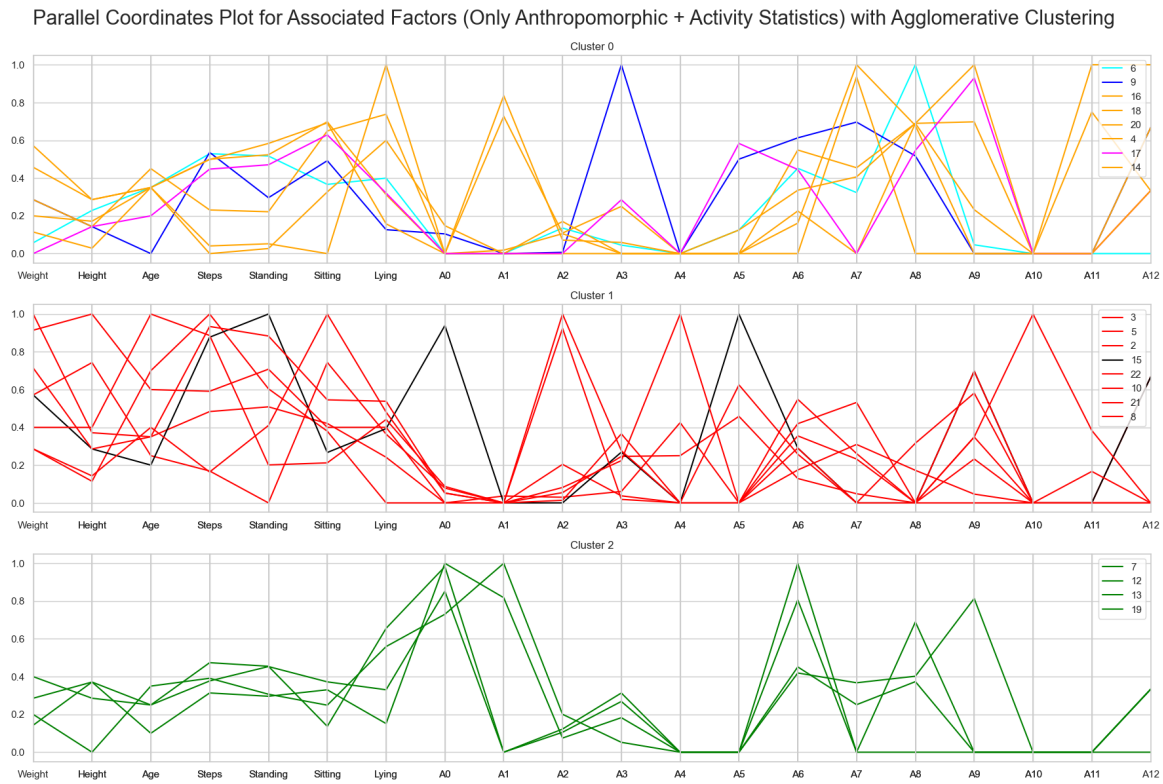

Figure 4.22: Parallel Coordinates plot of the last configuration for Associated Factors with the AC predictions

label. Thus, his associated factors do not justify his high PSQI label. Next, all features prior to A0 seem to follow a good sleep quality model for user 9. The next three dimensions are close to the poor sleepers, followed by six characteristics closer to people with a medium PSQI label, before finishing with four features like Cluster's 2. We could say that this subject shares a mixture of styles from each cluster; however, the majority of them are not close to Cluster 1. Regarding user 15, we see a pattern similar to the people with poor sleep quality for the first seven dimensions and from A5 to A11. A0 to A4 is closer to good sleep quality, while the last feature is more like a medium's PSQI label participant. Clearly, this user has more related factors with good or poor sleep quality than medium sleep quality. Ultimately, with the exception of anthropomorphic, A3, A6, and A7 features, user 17 shares all the rest dimensions with the yellow-colored cluster. The exceptions, though, are pretty close to subjects with low PSQI labels. These observations make sense since this user reported a good night's sleep, and the differences with a medium PSQI label characteristics are minor.

It is easily comprehensible that there is an association between daily activities and sleep quality. That is something we showed in the previous section too. Also, we can observe that some users' answers in the Pittsburgh Sleep Quality Questionnaire do not follow up with our findings. To enhance this theory, we present a plot of the PSQI of each user against their cortisol after sleep (Figure 4.23).

As stated in [4] and [20], sleep quality and the awakening cortisol, sampled by the saliva, have a negative relation. In our case, we can notice that the blue line indicates a positive relationship between these two variables. We notice that the average cortisol level of people with poor sleep quality is higher than that of participants with low PSQI values, which is abnormal. Definitely, user 2 also alters this result as even if he had a poor night's sleep, his cortisol level is the highest in our dataset; that makes him act like an outlier. As we mentioned earlier, we have evidence that users 9 and 15 probably did not have the sleep quality they reported; user 9 with high cortisol had a better sleep quality, while user 15 with low cortisol had a night of poor sleep. Having two different PSQI values for these participants would alter the gradient of the blue line towards the theory of the previously published paper.

There can be only two possible explanations for this phenomenon. The first is that there was an error during the laboratory analysis of the samples, which is unlikely. The second and most probable is that some participants provided mismatching answers with their actual sleep quality, either deliberately or subconsciously.

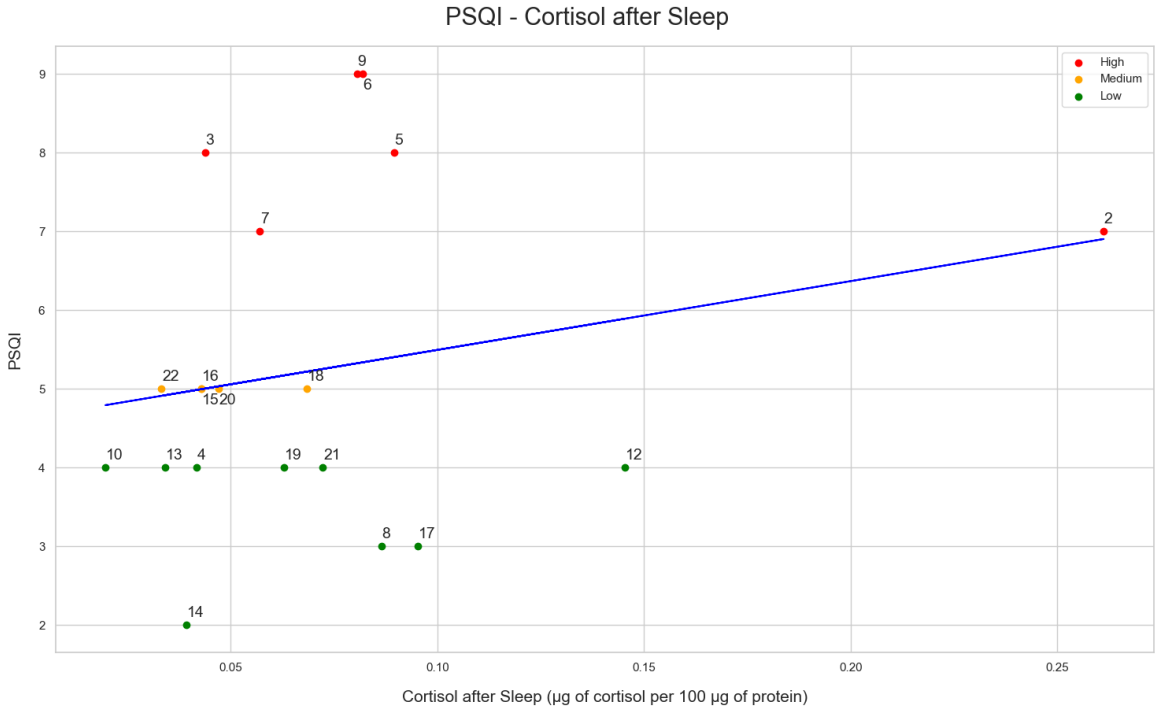

Figure 4.23: PSQI - Cortisol after sleep

4.5 Summary

To summarize our results chapters, we highlight our main findings. Regarding the HR patterns, we discovered that people with good sleep quality have a constant HR close to the average; people with PSQI equal to five show a very unstable HR; participants with high PSQI label demonstrate more frequent peaks in their HR with the mean of their normal HR being over the average HR of all users. Next, concerning daily external factors, we noticed a negative correlation between sleep quality and caffeinated drink consumption, sitting, and age. A positive correlation was observed with slow/medium walking, height, and the number of steps. There were also indications that heavy movement, negative feelings before bed, ans small screen time also affects it negatively. Finally, we detected that the data of users 9 and 15 suggest a sleep quality different from the one they reported. We accompanied that assumption by plotting a PSQI - Cortisol after sleep plot, which did not follow the theory, meaning that there is either a fault in the participants' answers or the cortisol measurements. That was everything about the results and our comments on them. Following, we are analyzing our proposals for further work, and eventually, we present the conclusions of this project.

Chapter 5

Further Work

We present in this chapter our thoughts of how this project could possibly be extended. Some of these suggestions were initially planned to be executed by us, but due to the limited time, we did not have the chance to do so. It is divided into two sections: (a) different techniques; (b) different features. The former is about working with the same features from the original dataset but using them with different parameters or algorithms. Concerning the second, we suggest additional features from the original dataset that could be used for the analysis - either with the same or different methodology.

5.1 Different Techniques

During our discussion of the results in the previous chapter, we noted some improvements that can take place. We analyze them sequentially.

First of all, we begin with the distinctive time series statistics. We witnessed there a massive increase in the rand score after we combined two feature selection methods. Thus, by extending the already applied methods or selecting new ones, there is a high possibility of improving the results even more. A couple of these methods can be the Multi-Cluster Feature Selection (MCFS) and the weighted k-means (W-k-means) [2]. The first [13] is a method that selects the best features to maintain the data's multi-cluster structure. W-k-means [26] is an alternation of the original k-means, which, additionally, automatically computes the weights of the features based on the variance of the within-cluster distances.

We have three thoughts regarding TSLearn's KMeans method. One, of course, is the configuration of the parameters. However, because of the total time required to finish the clustering process, we created a function that takes a reset indexed DataFrame along with the desired time window in seconds. It returns a new time series dataset that has one observation every desired_time_window seconds by taking the mean of each value for that period. Using the windowed DataFrames to the TSLearn's KMeans will make it run faster and alter the results. Finally, we realized how powerful motifs, discords, and snippets were. Hence, we thought of combining the two methods. It would be interesting to feed these extracted features in this method with the "dtw" as a metric.

The remaining extracted features are motifs, discords, and snippets. In this paragraph, we discuss some techniques that can be applied to all of them. As noted in the previous chapter, one method is to extract the K best features for every user. K is a positive integer provided by the researcher. For example, extract the best three discords from all users. That would result in an n x m x K matrix, where n is the number of users, m the length of the extracted subsequence, and K the number of extracted features for each user. Applying clustering techniques to this matrix would be something worth examining. Furthermore, exploring these features with various parameter configurations is another matter that can be investigated. Moreover, inspired by the DTW theory, we could find the best feature of each time series, then extract the top 10 occurrences, and finally, apply DBA [51] to these occurrences. The resulting subsequence can then be used as the feature of that user in the clustering algorithms.

Regarding motifs, there is a technique that can be used, named consensus motifs [30]. It is an approach that helps find similar motifs between the time series of all users. That would return motifs with minor differences between them, and therefore, different clustering results.

Additionally, there are more feature extraction methods that can be applied in our time series—for instance, the SAX [36] algorithm, which transforms the time series into symbolic representations. We

used one partitioning (KMeans) and one hierarchical (AC) clustering algorithm in our thesis. That means that different types of clustering algorithms can also be used for the corresponding task. An example is a density-based clustering algorithm like Kernal DBSCAN [14].

Subsequently, regarding the associated factors, we perceived how the rand score improved when we changed the features we used. Hence, applying feature selection algorithms in these dimensions is something that should be explored in the future.

Following, we present our suggestions based on other features that could be used as an extension to this research.

5.2 Different Features

It is apparent that the dataset from the original paper is rich in terms of features. We did not use three columns: Axis1, Axis2, and Axis3, which correspond to x, y, and z of the accelerometer data, respectively. We explained the reason why we did not use them to cluster sleep quality. However, there are researches like [44] that use this information to classify sleep quality. Furthermore, the accelerometer measurements prior to the sleep of the participants can also be used for research on sleep quality with techniques like the ones we applied in this project. On the same wavelength, HR data can also play a vital role in similar research, and consequently, associate daily HR with sleep quality.

Last but not least, the cortisol and melatonin measurements are essential indicators of sleep quality. Experimentation to assess how participants define their sleep quality and what their body measurements designate is something that can be done in order to evaluate the assumptions we made at the end of section 4.4. Some examples are the following: [4], which negatively correlates PSQI with the awakening cortisol, [24] that expresses the relation of melatonin and cortisol before and after sleep, and finally [20], which agrees with the previous statements and also appends a relation between melatonin and sleep quality.

Other than the related future work we mentioned above, the number of features of the original dataset is so significant that they can be used in numerous different projects. A simple example is how daily activities are affected by the participant's personality and psychological status.

The next chapter, which is also the last, is dedicated to analyzing the conclusions of this thesis.

Chapter 6

Conclusions

The final chapter in this dissertation is devoted to our resolutions. We witnessed how our alternative approach returned promising results. However, due to time restrictions, we did not have the chance to explore more features or techniques during our research. Many exciting new studies can be contacted, as we mentioned in the previous chapter. Furthermore, there is the presence of sample bias in our dataset, yet, our results complied with the ones of other studies as we observed. Following, we divide this chapter into three sections: (a) HR patterns and their association to sleep quality; (b) daily external factors linked to sleep quality; (c) outcome by combining the previous results.

6.1 Heart Rate Patterns of Sleep Quality

One of the two principal aims of this thesis was to find patterns in the HR of the participants during their sleep and link them to their sleep quality. We used several feature extraction methods, but the most promising and informative were derived from the matrix profile theory: the motifs and the snippets. We also created a clustering algorithm, called MyClusterAlgorithm, that takes equal-sized time series and divides them into three clusters. As discussed in Chapter 4, there are some indications that relate HR to sleep quality.

We begin by analyzing the relation between HR and low PSQI. We found that participants with good sleep quality have a constant HR close to the average HR of all users for more extended periods than other users. Also, all of them show a sudden increase in their HR that lasts for 15 to 20 seconds and then decays fast (Figures 4.6 and 4.7).

Following, people with poor sleep quality exhibit an increased constant HR compared to all users, something that is in agree with [11]. Moreover, these users show more frequent spikes in their HR compared to people with better sleep qualities. These HR rises last longer than people with good sleep quality, while the same applies to their decreases. Also, the HR reaches local maxima a lot quicker than people with other PSQI labels.

Lastly, participants with a PSQI equal to five showed a surprising fact. Contrarily to what a logical person would think, people with medium sleep quality had a very unstable HR for long-lasting periods, something that graphs of people with worse sleep quality did not indicate. Also, their average HR is below the respective value of all users. Last but not least, these users show a slower decay rate in their HR after a peak.

These were our conclusions about HR and how it is connected to sleep quality. Next, we present our findings on external factors and their relationship to sleep quality.

6.2 Associated Factors of Sleep Quality

The second major aim of this thesis was to analyze daily external factors and associate them to sleep quality. We achieved that by extracting the activity statistics, each user had reported during the experiment. We found that sleep quality improves when people drink less caffeine and sit for smaller periods, including studying, eating, or driving. Age also plays an important role in sleep quality since the results suggest that older people have a better night's sleep than younger. The opposite applies to taller people,

as poor sleep quality seems to be positively correlated with height. Other factors that also worsen sleep quality are light movement, like walking or working, and taking many steps during the day. Also, there were indications that heavy movement (e.g., gym), negative feelings prior to bed time, and small screen usage (e.g., mobile) affects sleep quality negatively.

These conclusions seem to follow what other researchers suggested too. [58] proposed that longer traveled distances are related to high PSQI, similar to our Steps feature. Also, [6] stated that sitting for a considerable amount of time negatively affects sleep quality, something that we mentioned too.

6.3 Combination of Heart Rate and Associated Factors

During our research, we combined the outcomes of HR clustering with the associated factors clustering. The examination of the combination that returned the best rand score, meaning best match, agreed on four users having different sleep quality than the one they reported in the questionnaire. Our analysis showed that we have a good reason to believe in our results since at least two of these participants strongly share features with people that showed a different sleep quality than theirs. We strengthened our assumption by plotting a PSQI against Cortisol after sleep plot, which did not follow the findings of previously published papers.

Bibliography

[1] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy. Adaptive mobile activity recognition system with evolving data streams. *Neurocomputing*, 150:304–317, 2015.

[2] S. Alelyani, J. Tang, and H. Liu. Feature selection for clustering: A review. *Data Clustering*, pages 29–60, 2018.

[3] P. J. M. Ali, R. H. Faraj, and E. Koya. Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1):1–6, 2014.

[4] J. Backhaus, K. Junghanns, and F. Hohagen. Sleep disturbances are correlated with decreased morning awakening salivary cortisol. *Psychoneuroendocrinology*, 29(9):1184–1191, 2004.

[5] E. Baehr, W. Revelle, and C. Eastman. Individual differences in the phase and amplitude of the human circadian temperature rhythm: with an emphasis on morningness–eveningness. *Journal of Sleep Research*, 9:117 – 127, 06 2000.

[6] Y. Bai, B. Xu, Y. Ma, G. Sun, and Y. Zhao. Will you have a good sleep tonight? sleep quality prediction with mobile phone. In *BODYNETS 2012 - 7th International Conference on Body Area Networks*, pages 124–130, 2012.

[7] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*, pages 1–17. Springer, 2004.

[8] J. E. Beck and B. P. Woolf. High-level student modeling with machine learning. In *International Conference on Intelligent Tutoring Systems*, pages 584–593. Springer, 2000.

[9] A. Bhandari. Feature scaling for machine learning: Understanding the difference between normalization vs. standardization. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/#h2_9, 2020. (Accessed: 21.06.2021).

[10] J. Brownlee. Logistic regression for machine learning. https://machinelearningmastery.com/logistic-regression-for-machine-learning/, 2020. (Accessed: 09.09.2021).

[11] A. R. Burton, K. Rahman, Y. Kadota, A. Lloyd, and U. Vollmer-Conna. Reduced heart rate variability predicts poor sleep quality in a case–control study of chronic fatigue syndrome. *Experimental brain research*, 204(1):71–78, 2010.

[12] Zhenyu C., M. Lin, C. Fanglin, N. D. Lane, G. Cardone, W. Rui, L. Tianxing, C. Yiqiang, T. Choudhury, and A. T. Campbell. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, pages 145–152, 2013.

[13] D. Cai, C. Zhang, and X. He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 333–342, New York, NY, USA, 2010. Association for Computing Machinery.

[14] S. Chandrakala and C. C. Sekhar. A density based method for multivariate time series clustering in kernel feature space. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1885–1890. IEEE, 2008.

[15] J. Chen, Y. Liu, M. Gao, Z. He, and Z. Yu. Battery charging and discharging feature extraction method based on the best u-shapelets. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pages 207–211. IEEE, 2018.

[16] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018.

[17] G. Curcio, D. Tempesta, S. Scarlata, C. Marzano, F. Moroni, P. Rossini, M. Ferrara, and L. De Gennaro. Validity of the italian version of the pittsburgh sleep quality index (psqi). *Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 34, 04 2012.

[18] D. De Paepe, S. Vanden Hautte, B. Bram Steenwinckel, F. De Turck, F. Ongenae, O. Janssens, and S. Van Hoecke. A generalized matrix profile framework with support for contextual series analysis. *Engineering Applications of Artificial Intelligence*, 90:103487, 2020.

[19] O. T. S. Deviations. An overview of correlation measures between categorical and continuous variables. https://tinyurl.com/3tjdrvx6, 2018. (Accessed: 09.09.2021).

[20] A. Dubberke, N. Falkenreck, F. Nisius, E. Ellsiepen, and J. Hellhammer. Impact of morning cortisol and evening melatonin secretion on sleep quality. https://stresszentrum-trier.de/fileadmin/docs/stresszentrum-trier/img/PDF/2019_Poster_Cort-Sleep-Melatonin_WASAD.pdf, 2019. (Accessed: 11.09.2021).

[21] M. Ferrara and De Gennaro L. How much sleep do we need? *Sleep Medicine Reviews*, 5:155–179, 2001.

[22] H. Fritz, L. A. García-Escudero, and A. Mayo-Iscar. A fast algorithm for robust constrained clustering. *Computational Statistics & Data Analysis*, 61:124–136, 2013.

[23] A. Fu, O. Leung, E. Keogh, and J. Lin. Finding time series discords based on haar transform. In *International Conference on Advanced Data Mining and Applications*, pages 31–41. Springer, 08 2006.

[24] M. Garaulet, P. Gómez-Abellán, and J. A. Madrid. Chronobiology and obesity: the orchestra out of tune. *Clinical Lipidology*, 5(2):181–188, 2010.

[25] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. *Data Mining and Knowledge Discovery*, 34:1104–1135, 2020.

[26] J.Z. Huang, M. K. Ng, R. Hongqiang, and L. Zichen. Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):657–668, 2005.

[27] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[28] S. Imani, F. Madrid, W. Ding, S. Crouter, and E. Keogh. Matrix profile xiii: Time series snippets: A new primitive for time series data mining. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, pages 382–389, 2018.

[29] M. I Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[30] K. Kamgar, S. Gharghabi, and E. Keogh. Matrix profile xv: Exploiting time series consensus motifs to find structure in time series sets. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1156–1161, 2019.

[31] O. M. Lateef and M. O. Akintubosun. Sleep and reproductive health. *Journal of circadian rhythms*, 18:1–1, 2020.

[32] D. S. Lauderdale, J. H. Chen, L. M. Kurina, L. J. Waite, and R. A. Thisted. Sleep duration and health among older adults: associations vary by how sleep is measured. *Journal of epidemiology and community health*, 70:361–366, 2016.

[33] Sean M. Law. STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining. *The Journal of Open Source Software*, 4(39):1504, 2019.

[34] W. Lee and G. C. Migliaccio. Field use of physiological status monitoring (psm) to identify construction workers' physiologically acceptable bounds and heart rate zones. In *2014 International Conference On Computing in Civil and Building Engineering*, pages 1037–1044, 2014.

[35] J. Lim, S. Chung, K. J. Noh, G. G. Kim, and H. T. Jeong. An empirical study on finding experience sampling parameters to explain sleep quality based on dimension reduction. In *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1295–1299, 2019.

[36] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, page 2–11, New York, NY, USA, 2003. Association for Computing Machinery.

[37] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo. Daily activity feature selection in smart homes based on pearson correlation coefficient. *Neural Processing Letters*, 51(2):1771–1787, Apr 2020.

[38] L. Luo and S. Lv. An accelerated u-shapelet time series clustering method with lsh. In *Journal of Physics: Conference Series*, volume 1631, page 012077. IOP Publishing, 2020.

[39] Z. Ma, C. Chen, M. Wang, Y. Zhao, L. Ying, G. Wang, and J. Zhao. A low-power heart rate sensor with adaptive heartbeat locked loop. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021.

[40] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[41] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008.

[42] W. McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[43] Q. Meng and P. Pu. Rls: An efficient time series clustering method based on u-shapelets. *Intelligent Data Analysis*, 22(4):767–785, 2018.

[44] H. Miwa, S. I. Sasahara, and T. Matsui. Roll-over detection and sleep quality measurement using a wearable sensor. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1507–1510, 2007.

[45] I. Mohamad and D. Usman. Standardization and its effects on k-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, 6:3299–3303, 09 2013.

[46] D. Morelli, A. Rossi, L. Bartoloni, M. Cairo, and D. A. Clifton. Sdnn24 estimation from semi-continuous hr measures. *Sensors*, 21:1463, 2021.

[47] F. Murtagh and P. Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

[48] P. Patel, E. Keogh, J. Lin, and S. Lonardi. Mining motifs in massive time series databases. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 370–377, 2002.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] I. Perez-Pozuelo, M. Posa, D. Spathis, K. Westgate, N. Wareham, C. Mascolo, S. Brage, and J. Palotti. Detecting sleep in free-living conditions without sleep-diaries: a device-agnostic, wearable heart rate sensing approach. *medRxiv*, pages 2020–09, 2021.

[51] F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3):678–693, 2011.

[52] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[53] T. Roenneberg. The human sleep project. *Nature*, 498:427–428, 2013.

[54] L. Rokach and O. Maimon. Clustering methods. In *Data mining and knowledge discovery handbook*, pages 321–352. Springer, 2005.

[55] A. Rossi, E. Da Pozzo, D. Menicagli, C. Tremolanti, C. Priami, A. Sîrbu, D. Clifton, C. Martini, and D. Morelli. A public dataset of 24-h multi-levels psycho-physiological responses in young healthy adults. *Data*, 5:91, 2020.

[56] A. Rossi, D. Pedreschi, D. Clifton, and D. Morelli. Error estimation of ultra-short heart rate variability parameters: Effect of missing data caused by motion artifacts. *Sensors*, 20, 2020.

[57] S. Sahin, K. Ozdemir, A. Unsal, and N. Temiz. Evaluation of mobile phone addiction level and sleep quality in university students. *Pakistan journal of medical sciences*, 29(4):913–918, Jul 2013.

[58] A. Sano, A. J. Phillips, A. Z. Yu, A. W. McHill, S. Taylor, N. Jaques, Charles A. C., E. B. Klerman, and R. W. Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6, 2015.

[59] A. Sathyanarayana, J. Srivastava, and L. Fernandez-Luque. The science of sweet dreams: Predicting sleep efficiency from wearable device data. *Computer*, 50(3):30–38, 2017.

[60] J. Starkweather and A. K. Moske. Multinomial logistic regression. http://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf, 2011. (Accessed: 09.09.2021).

[61] I. A. Szöke, V. Stoicu-Tivadar, and D. Lungeanu. Sleep fragmentation. a study on how daily activities affect our sleeping. In *2015 IEEE 19th International Conference on Intelligent Engineering Systems (INES)*, pages 259–263, 2015.

[62] R. Tavenard, J. Faouzi, G. Vandewiele, Felix Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020.

[63] S. Taylor. Introduction to time series data and analysis. https://soft-dev.org/events/bench16/slides/Simon_Taylor.pdf, 2016. (Accessed: 21.06.2021).

[64] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[65] Wikipedia contributors. Logistic function — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Logistic_function&oldid=1043138185, 2021. (Accessed: 09.09.2021).

[66] L. Ye and E. Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 947–956, 2009.

[67] C. C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. Keogh. Matrix profile i: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1317–1322, 2016.

[68] J. Zakaria, A. Mueen, and E. Keogh. Clustering time series using unsupervised-shapelets. In *2012 IEEE 12th International Conference on Data Mining*, pages 785–794, 2012.

[69] Wenyao Zhu. *Time-Series Feature Extraction in Embedded Sensor Processing System*. PhD thesis, KTH Royal Institute Of Technology, 2020.

[70] Y. Zhu, A. Mueen, and E. Keogh. Matrix profile ix: Admissible time series motif discovery with missing data. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2616–2626, 2021.

[71] Y. Zhu, Z. Zimmerman, N. S. Senobari, C. C. M. Yeh, G. Funning, A. Mueen, P. Brisk, and E. Keogh. Matrix profile ii: Exploiting a novel algorithm and gpus to break the one hundred million barrier for time series motifs and joins. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 739–748, 2016.

[72] P. Zoumpoulidis. Exploring multilevel monitoring for sleep quality and its associated factors. https://github.com/zoump/thesis, 2021. (Accessed: 13.09.2021).

[73] P. Zoumpoulidis. Exploring multilevel monitoring for sleep quality and its associated factors - presentation. https://www.youtube.com/watch?v=BLmReydaFEo, 2021. (Accessed: 17.09.2021).

Appendix A

Clustering Results

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Extracted Stats with KMeans	0.4947	1	0	1	1	2	1	2	0	1	1	1	2	2	1	1	0	2	1	1	1
Extracted Stats with AC	0.4947	0	1	0	0	2	0	2	1	0	0	0	2	2	0	0	1	2	0	0	0
Extracted Stats with Feature Selection and KMeans	0.6105	1	0	0	1	1	0	0	1	2	0	1	0	0	0	0	0	0	0	0	0
Extracted Stats with Feature Selection and AC	0.5789	2	0	0	2	1	0	0	2	0	1	2	0	1	0	0	0	0	0	0	0
TSLearn KMeans with DTW	0.5158	0	1	1	1	1	0	1	2	0	1	1	1	1	1	1	1	1	1	1	1
Motifs with KMeans and 90s Length	0.6526	1	1	1	0	2	0	0	1	1	2	2	2	2	2	2	2	2	2	1	2
Motifs with AC and 90s Length	0.6526	0	0	0	2	1	2	2	0	0	1	1	1	1	1	1	1	1	1	0	1
Motifs with MyClusterAlgorithm and 90s Length	0.6263	1	1	1	2	2	2	2	1	1	0	0	0	0	0	2	2	0	0	1	0
Motifs with KMeans and 120s Length	0.5000	2	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
Motifs with AC and 120s Length	0.5474	0	2	1	0	2	0	0	1	2	2	2	2	0	2	2	2	2	2	1	1
Motifs with MyClusterAlgorithm and 120s Length	0.5368	2	1	2	1	1	0	1	2	1	1	0	1	0	0	1	1	0	0	2	2
Motifs with KMeans and 300s Length	0.5053	0	1	1	0	1	0	1	1	2	1	1	1	1	1	1	1	1	0	0	0
Motifs with AC and 300s Length	0.4632	1	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	1	0	0
Motifs with MyClusterAlgorithm and 300s Length	0.5263	0	1	0	2	0	0	0	1	2	0	2	0	0	0	0	0	0	0	2	1
Motifs with KMeans and 600s Length	0.5789	0	1	2	0	1	2	2	1	2	1	1	1	2	2	2	2	2	1	2	2
Motifs with AC and 600s Length	0.6632	1	0	2	1	0	2	0	0	0	0	0	0	2	2	2	2	2	0	2	2
Motifs with MyClusterAlgorithm and 600s Length	0.5684	0	2	1	2	1	0	2	0	2	0	2	0	1	0	1	0	0	2	1	0
Discords with KMeans and 90s Length	0.4947	0	0	2	0	2	0	2	0	0	2	2	0	0	2	0	2	2	1	0	0
Discords with AC and 90s Length	0.4316	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
Discords with MyClusterAlgorithm and 90s Length	0.5684	1	1	1	1	0	0	1	2	2	0	2	2	2	0	1	1	1	0	1	0
Discords with KMeans and 120s Length	0.5158	0	2	2	0	2	0	2	2	1	2	2	2	2	2	2	2	2	0	2	2
Discords with AC and 120s Length	0.4632	0	2	2	0	2	2	2	2	1	2	2	2	2	2	2	2	2	0	2	2
Discords with MyClusterAlgorithm and 120s Length	0.5947	1	2	1	1	2	1	1	0	0	2	0	0	2	0	0	1	0	1	0	1
Discords with KMeans and 300s Length	0.5263	0	2	1	0	1	2	2	2	2	1	1	2	2	1	2	1	2	0	1	2
Discords with AC and 300s Length	0.5263	0	2	1	0	1	2	2	2	2	1	1	2	2	1	2	1	2	0	1	2
Discords with MyClusterAlgorithm and 300s Length	0.5263	0	2	2	2	0	1	2	0	0	0	0	1	0	0	2	0	0	1	0	2
Discords with KMeans and 600s Length	0.5158	2	2	0	2	0	2	2	0	0	0	0	0	0	2	2	0	0	1	0	2
Discords with AC and 600s Length	0.5158	1	1	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	2	0	1
Discords with MyClusterAlgorithm and 600s Length	0.5316	2	1	0	1	2	2	2	0	0	1	2	1	0	1	2	2	1	2	2	2
Snippets with KMeans and 90s Length	0.5368	0	1	1	1	2	1	1	2	1	2	2	2	0	1	1	2	2	0	1	2
Snippets with AC and 90s Length	0.5158	2	0	0	0	1	0	0	1	0	1	1	1	0	0	0	1	1	0	0	1
Snippets with MyClusterAlgorithm and 90s Length	0.5526	0	1	1	0	1	0	0	2	0	1	0	0	0	0	0	2	0	2	2	0
Snippets with KMeans and 120s Length	0.5947	0	1	1	0	1	2	2	2	0	1	2	2	0	2	0	2	2	2	0	2
Snippets with AC and 120s Length	0.6000	0	1	1	0	1	2	0	2	0	1	2	2	0	2	0	2	2	2	0	2
Snippets with MyClusterAlgorithm and 120s Length	0.5526	2	1	1	2	1	2	2	0	2	1	2	2	2	2	2	0	2	0	0	2
Snippets with KMeans and 300s Length	0.5474	2	2	2	1	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0
Snippets with AC and 300s Length	0.4842	0	0	0	1	2	0	0	0	0	2	2	0	0	0	0	2	0	0	2	0
Snippets with MyClusterAlgorithm and 300s Length	0.6105	2	1	0	2	1	2	0	1	0	1	0	2	2	2	2	1	0	2	1	2
Snippets with KMeans and 600s Length	0.5158	1	0	0	1	2	0	0	0	1	0	0	1	1	0	1	0	0	1	1	0
Snippets with AC and 600s Length	0.4947	1	0	0	1	2	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
Snippets with MyClusterAlgorithm and 600s Length	0.6211	1	0	2	1	1	0	0	2	2	2	2	1	1	1	1	2	0	1	2	1

Figure A.1: Sleep Quality Clustering Results

Method	Rand Score	6 - H	9 - H	3 - H	5 - H	2 - H	7 - H	15 - M	16 - M	18 - M	20 - M	22 - M	4 - L	10 - L	12 - L	13 - L	19 - L	21 - L	8 - L	17 - L	14 - L
Ass. Factors Using Questionnaire Results with KMeans	0.4526	0	0	0	0	0	1	0	0	1	0	0	0	2	1	0	1	0	0	0	1
Ass. Factors Using Questionnaire Results with AC	0.5789	2	2	2	1	2	0	1	2	0	1	1	2	1	0	0	0	2	2	2	0
Ass. Factors Using Questionnaire Labels with KMeans	0.5053	1	1	2	0	0	1	0	0	1	1	1	2	1	0	0	0	1	1	1	2
Ass. Factors Using Questionnaire Labels with AC	0.5316	0	0	1	0	2	0	2	2	0	0	0	1	0	2	2	2	0	1	0	1
Ass. Factors Using Only Activity Stats with KMeans	0.5895	0	0	2	1	1	0	2	2	1	2	1	0	1	0	0	0	2	2	2	0
Ass. Factors Using Only Activity Stats with AC	0.5316	0	0	1	1	1	2	1	0	0	0	1	0	1	2	2	2	1	1	0	0

Figure A.2: Associated Factors Clustering Results

Appendix B

Execution Instructions

The whole code, along with the accompanying data, are bundled together and available on GitHub [72]. Note that Python 3.9 is required to run. All libraries that need to be installed are inside the libraries.pdf file.

If the code is running locally, only the main.py needs to be executed. Otherwise, if the code is running online (like Google Colaboratory), the path from files createNeededDirectoriesIfNotExist and getUserFolders needs to change to the correct ones. Afterward, only main.py needs to be executed. The results after the execution of the code will be in folders DataFrames and Plots.