Bachelor Project

Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Digital Phenotyping Analysis

Michal Kubina

Supervisor: doc. Ing. Daniel Novák, Ph.D.
Field of study: Open Informatics
Subfield: Information and Computer Sciences
May 2021

Acknowledgements

I would like to express my particular gratitude to my supervisor, doc. Ing. Daniel Novák, Ph.D., for his guidance and constant support during a whole year of preparation for this bachelor thesis. Moreover, I am thankful to Ing. Jakub Schneider for his valuable inputs. Lastly, I would like to thank my family, which gave me the opportunity to attend university and which shows me constant support.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, 19. 5. 2021

Michal Kubina

Abstract

The main goal of this bachelor thesis is to examine digital phenotyping and provide an analysis of this phenomenon to evaluate the importance of digital phenotyping for future medical use.

The first part of this thesis describes the theoretical part focusing on people's data collected from smart devices, mainly from smartphones, and simultaneously related computer science focusing on mental health using these data is introduced. Consequently, thorough concepts and terms used in this area are explained, and state of the art is described.

Further, data collection platforms for smartphones are described and detailly compared by running a small pilot study. Then another pilot study is deployed to collect passive and active data from smartphones and actigraphy data from the wristband. These data entries are evaluated, visualised and assessed. Moreover, the GPS sensor, power state and actigraphy data are preprocessed, and meaningful features are calculated using these data types.

The second part focuses on the exploratory analysis of the data set gathered in the pilot study. The data are assessed as a whole unit and simultaneously individually by a person. Moreover, the classification of people's emotional states developed from the circumplex model of affect is introduced. This classification is done using a random forest algorithm based on the calculated features, and its ultimate goal is to distinguish between a person's emotional state.

Keywords: digital phenotyping, actigraphy, classification, behavioural analysis

Supervisor: doc. Ing. Daniel Novák, Ph.D.

Abstrakt

Hlavním cílem této práce je evaluace a analýza digitální fenotypizace, jakožto nového fenoménu, který by mohl být vhodný pro zdravotní péči budoucnosti.

První část této bakalářské práce popisuje sběr dat z mobilních zařízení a zároveň představuje odvětví informatiky, které se zabývá nemocemi duševního zdraví. Zároveň jsou zde popsány a vysvětleny základní koncepty a pojmy v rámci současného pokroku založené na vědeckých článcích a studiu odborné literatury.

Dále jsou v této bakalářské práci popsány platformy, které jsou určené ke sběru dat z mobilních zařízení. Tyto platformy byly vyzkoušeny v rámci pilotní studie a poté také srovnány odpovídajícím způsobem. Na základě výsledků této studie byla spuštěna další pilotní studie, ve které se sbírala aktivní a pasivní data a dále také aktigrafická data z chytrého náramku. Tato práce se ale zejména zaměřuje na data z GPS sensoru, četnost používání mobilu a data z chytrého náramku. Na základě těchto datových typů je spočítána řada příznaků.

Druhá část této bakalářské práce se zaměřuje na explorační analýzu dat sesbíraných v již zmíněné pilotní studii. Tato data jsou vizualizována a analyzována ze dvou úhlů pohledu. Prvním je pohled, ve kterém se data berou jakožto celek, přičemž druhý zkoumá data jednotlivců. Dále je zde představeno pojmenování lidských emocí na základě kružnicového modelu afektu. Tyto lidské emoce jsou pak klasifikovány náhodným lesem na základě již spočítaných příznaků. Konečný cíl klasifikátoru je rozeznat emoce lidí v rámci jednotlivých dní.

Klíčová slova: digitální fenotypizace, aktigrafie, klasifikace, behaviorální analýza

Překlad názvu: Behaviorální analýza dat

Contents

Figures

Tables

Chapter 1

Introduction

Nearly every person in modern society owns a smartphone connected to the internet network. Even in developing countries, this trend is growing day by day at an immense speed. Moreover, companies collect a vast amount of data from these devices, sometimes without users' explicit knowledge. At the same time, people are more prone to suffer from mental health illnesses than ever before and simultaneously, the cost of medical care is becoming higher and inaccessible for a large portion of the population. Here I see significant potential by connecting digital phenotyping and medical care for mental health problems.

Thus, in this thesis I focus on exploring digital phenotyping as a phenomenon. Precisely, I focus on the data that can be acquired through smartphone and the available data collection platforms. Moreover, I compare these complete solutions by running a small pilot study.

Consequently, another pilot study is deployed, focusing on data collection between twenty healthy participants using the Beiwe platform for two months. As the next step, actigraphy as another phenomenon is introduced, and participants of the study wear corresponding wrist bands collecting actigraphy data.

In the next phase, I preprocess data and calculate meaningful features. Then I apply an exploratory data analysis to this data set by considering the data set as a whole unit and simultaneously individually by a person. Moreover, I implement the circumplex model of affect, and thus people's mood is categorised by their emotional states.

Lastly, I apply a random forest classification of the emotional states. The data set is split into the train and test set for this classification. On the train set upsampling method is applied due to an imbalance of categories, and the repeated 10-fold cross-validation is done for tuning hyperparameters using grid search. The results are compared and elaborated on from different points of views.

To sum up, I consider this thesis a detailed introduction to digital phenotyping, a relatively unexplored and rising topic with considerable potential.

Chapter 2

Digital Phenotyping

2.1 Significance and Meaning

Nearly everybody owns a smart device such as a mobile phone, and people are starting to wear watches with various functions more and more. In the United States of America in 2015, 64% of the adult population owned a smartphone, and there is a likelihood that this number will increase [SC15]. In 2018, around 4.3 billion people were using the mobile internet network, and by 2024 the number is expected to reach 6.2 billion users [Cen19]. However, the development of these devices' full scale usage might be a little bit stagnating. With multiple sensors these devices have and the data they provide, there comes a massive opportunity for not only health science [MW18].

A relatively new phenomenon, called *digital phenotyping*, was first defined in 2016 as a "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" [TKLO16]. In other words, with digital phenotyping, we can collect several types of real-time data from a smart device and then deduce, classify or predict behaviour on our sample.

The data can be divided into two groups. *Active data* need intended engagement from an individual. Typically, it can be an answer to a survey. On the other hand, *passive data* do not require active participation and are collected in the background.

2.2 State of The Art

2.2.1 Description

The use of digital phenotyping can be found mainly in the health industry. Moreover, it is still being perceived and developed as a clinical tool for psychiatric patients [Ins18]. The amount of data we can collect about a person with a smartphone might be considered enormous and simultaneously allows us to access people's lives in a great deal. A significant advantage is the constant monitoring of psychiatric patients' activity compared to a few doctor's visits per year [OR16]. Furthermore, the ultimate goal which lies ahead of us is to preempt mental health illnesses and intervene in real time using behavioural analysis and the collected data [OR16].

However, until 2020 no algorithm approved by governmental regulatory agencies for psychological or psychiatric use was allowed to predict a clinical event [CSL+20]. Thus, the use of digital phenotyping is still in its infancy.

The reason might be that technology is more advanced year by year, but health science does not cooperate significantly with these developments. Simultaneously, there is a lack of methods to analyse and come to a conclusion from the unique and intensive data sets [MW18]. As proof, from 2004 to 2017, there was a growing trend in the number of science articles regarding digital phenotyping [LXZ19]. However, in 2017 the number of studies published on this topic was just around 350 [LXZ19].

Though, this does not mean that digital phenotyping is not used. There were several pilot studies made focusing on psychiatric patients.

2.2.2 Bipolar Disorder

The usage of digital phenotyping in studies focusing on bipolar patients and their diagnoses, which aims to achieve classification or prediction of relapses in the future, seems to be getting more and more popular. However, it appears to be an arduous task because bipolar patients are highly heterogeneous [BHI+19].

Some studies use the speech rhythm of incoming and outcoming calls to classify patients' relapses [MGMP17]. Other research uses derived speech from calls and the number of calls and text messages [FJBF+16].

Another pilot study focuses on the mobile's keyboard and typing frequency for identifying mood [CZZ+17]. Moreover, there is research on predicting relapses using a keyboard, which takes advantage of typing frequency and the length of typing sessions and the total usage of a keyboard per week [ZAR+17].

Moreover, sometimes GPS location or other passive data are used for detecting depression of bipolar patients [PTS+16]. Furthermore, some studies combine GPS and accelerometer sensor [GOB+14].

2.2.3 Another Usage

Even though bipolar disorder gets the most attention in digital phenotyping, some studies focus on relapse prediction in schizophrenia using passive data such as GPS, accelerometer, call and text message logs, phone charging status or screen/off time [BTS+18]. Moreover, the severity of social anxiety disorder was analysed by movement data and the number of calls [JSW20]. Passive data from phones were also used to track rigour and course of depression and conspicuous Internet usage [MBM+14].

Nonetheless, there is a potential for diseases that affect patients' physical state. As proof, another research focuses on tracking speech decline with amyotrophic lateral sclerosis patients using speech audio recording [CGP+19]. The related study tries to quantify this disease's progression using speech recordings and passive data from phone such as GPS location, accelerometer, WI-FI usage or communication metadata [BPC+19].

An exciting course to take in digital phenotyping might also be tracking people's social network behaviour. In a recent study [ESM+18], the textual content of Facebook posts and their length and frequency were used to predict depression.

2.3 Ethics

Digital phenotyping might sound like an ideal and promising method for the future of health science and the patients themselves. However, we must acknowledge that a smartphone can provide us with a crucial data set, including social, environmental, physiological or interpersonal details. These are unique pieces of information about a person's life that should remain confidential and definitely cannot be shared in any circumstances. For example, in a study [HTL19] regarding data and privacy policy within 36 most downloaded mobile applications focusing on mental health or smoking cessation apps, 80% shared their data with third-party commercial entities and just 40% of them disclosed this in their privacy policy.

We can track people's movements, sleep habits, routines, location, social network usage, and a nearly indefinite amount of other data with digital phenotyping. The privacy issue, should be taken seriously and with rigorous thought.

Firstly, it is necessary to disclose and explain data collection within the patients in order to gain their trust. A patient with a mental illness should not feel surveilled and under the threat of possible data breach. For instance, just four spatiotemporal points are needed to identify 95% of individuals who collect data hourly, resulting in 24 points per day [MHVB13].

Secondly, the data should be anonymised. To illustrate, this can be done through a unique identifier that can still distinguish a single person without disclosing the name. Moreover, the information and the data should be encrypted at all times and handled carefully [Onn20].

To sum up, digital phenotyping is a new scientific discipline with no specific standards for recording behaviour on phones, and there is no legislation regarding digital phenotyping, which could increase trust among the patients or professional public [MSH20].

2.4 Predecessors

Even though digital phenotyping is a relatively new phenomenon, there have always existed tendencies across the professional community to track human behaviour in order to cure mental illnesses. It all started with experience sampling methodology *(ESM)* created by Larson and Csikszentmihalyi in 1983 [LC14]. This approach is based on responding to questionnaires during the day at random times for multiple days. A similar idea, ecological momentary assessment *(EMA)*, can be described as a wide set of methods that collects experiences and current behaviours about a subject in its natural environment close to real-time [SSH08].

Historically, for *EMA/ESM* studies, which are based on self-reported questionnaires, paper formats surveys were used, then personal digital assistants without an internet connection, and eventually, devices with an internet connection [DIJ13].

However, when smartphones were invented, many people realised their potential to replace questionnaires and even use them in various ways to improve people's health. Consequently, mobile health applications became popular, and the *mHealth* industry was created [OR16].

Mobile health applications can be considered a platform for self-reporting information from nutrition to exercise or mental health. However, only recently, in 2015, these apps were used for immediate intervention [BZSB+14]. Thus, someone might consider digital phenotyping using a smartphone as a part of the *mHealth* area.

2.5 Challenges

Data collection in digital phenotyping unquestionably presents challenges that need to be overcome. Firstly, it is vital to reduce the amount of active data that might burden the participants who may withdraw from the projects or stop providing active data by neglecting their responsibilities [Onn20]. Furthermore, active data can be subjective, distorting analysis and results [TRN20].

Secondly, considering data collection through a smartphone, it is challenging and simultaneously crucial to stabilise the data sets where each phone might have different data characteristics, even from the same sensor types. Moreover, it is necessary to rely on raw data, as health SDK's, which provide analysed data and insight into a person's health, from two leading software providers, Apple and Google, cannot deliver what we need. [Onn20]

Furthermore, we should not be limited to a specific mobile platform, as there are differences in users' characteristics. As proof, iOS users have a higher annual income than Android users, $53,000$\$ to $37,000$\$, which might also reflect the mobile users' behaviour [New18].

Lastly, a smartphone is a complex device that relies on a battery. Thus, we might found extended periods of missing data, which might impair data sets. This problem might be caused by technical difficulties with sensors or user's intervention by uninstalling an app or turning some sensors off. For example, in 6 studies across 211 participants, it was found that the non-collection rate of the GPS sensor was 14%, and the accelerometer's non-collection proportion was 25% [KCK+21].

Nonetheless, as mentioned in Section 2.3, data and privacy policy issues, in general, need to be solved and taken into account when moving forward.

2.6 Conclusion

There exists a big hope across the professional community that digital phenotyping will be a tool for effectively tackling mental health disorders due to multiple reasons.

This mechanism allows for continuous monitoring of patients, which is especially valuable when they visit their doctor a few times per year. Also, individuals at heightened risk of suffering from mental disorders could be identified cost-effectively and on a broad scale. Most importantly, a retrospective analysis could provide a missing variable in preventing frequent events for individuals and infrequent events for the population. [OR16]

Furthermore, digital phenotyping allows us to see how the patient functions at home, not at the clinic, which is a crucial and never seen before advantage in the treatment [Ins18]. Another advantage might be that gathered empirical data might refine and advance progress in treating mental health problems, and simultaneously, it could allow just in time intervention for patients with mental illnesses [MW18].

Consequently, we could achieve enhanced mental health across the population [OR16].

Contrary, some opinions are not that optimistic about digital phenotyping. Firstly, there might be a significant risk of false positives, which essentially means misdiagnosing anomalies in a healthy community as a risk sign. Secondly, some mental health disorder patients deny that they have a problem. Thirdly, there is higher awareness about the usage of private data from businesses. Lastly, the ethical challenge of data privacy remains unanswered. [Tek20]

Chapter 3

Data Collection Platforms

3.1 Introduction

Data from digital phenotyping can also be described as an individual's behavioural data because it represents a person's behaviour and actions. However, collecting raw and quality data can be an arduous task. Especially when talking about information from smartphones.

Thus, when one wants to collect behavioural data from smartphones, unique collection platforms were developed which focus on this issue and can deliver raw data in high quality across the most important mobile operating systems and mobile phone brands. This chapter will describe available mobile platforms used across the professional community.

A data collection platform might be a vast and broad concept. Usually, it is a complete solution for gathering data on smartphones. Thus, it includes a backend server for data collection itself, a dashboard for researchers and a mobile application for a patient, which needs to be installed on a smartphone. [GPT+13]

Though one might suggest a surplus of these complete solutions, it is not valid. There are just two complete data collection platforms, open-source and available for use or pilot studies.

3.2 Beiwe

3.2.1 Description

A data collection platform for digital phenotyping called Beiwe is a solution that is innovative and was one of the first in the field. Beiwe was introduced in 2016 by the Onnela Lab at the Harvard TH Chan School of Public Health. This project aimed to create a complete solution with a study portal, database, data analysis and smartphone app for biomedical research focusing on raw and quality data, which would create unified datasets. The concept works but is still in development and is being improved continuously. [TKLO16]

The platform's backend server is designed to work on Amazon Web Services, so it is necessary to use these settings to create a study. However, this platform's core lies in the main dashboard accessed through a website, where single studies can be reached, created or modified by researchers and users with appropriate permissions. Moreover, the data can be downloaded through the dashboard.

A patient needs to install a corresponding application available for Android and iOS devices for free. In other words, the application, and thus the platform, is available for just Android and iOS devices.

A part of this project which could analyse and preprocess available data is currently deprecated, and by the time this thesis was being done, a new analysis package and preprocessing library were in development.

3.2.2 Data Collection

Beiwe collects raw data, trying to have a unified, durable and homogeneous data set across all mobile phones and available mobile operating systems. However, mobile platforms differ, and thus, it is impossible to have the same data types coming from the two most spread operating systems Android and iOS [O'D21]. The reason might be that their security and privacy policies differ, and the data available from different sensors is not the same.

In Table 3.1, we can find a list of passive data types that Beiwe platform captures on each mobile platform and alternatively mobile operating system if the data is platform-specific. Further, Beiwe advertises the possibility of collecting calls and text messages logs. Nonetheless, this is not currently possible as Apple's and Android's privacy policies changed, and Beiwe did not edit their policy utilisation yet.

The considerable advantage of data management by Beiwe is data scalability. We can set up different capturing frequency for each sensor mentioned in Table 3.1. This option gives us a possibility to change the sampling frequency of sensors according to the study needs, and with this way, we can optimise the data set size. Moreover, we can choose which specific sensors we want to use to collect data.

Other types of data the Beiwe platform provides are active data. This type of data needs actual engagement from a patient. In Table 3.2, types of information available to collect are listed. In general, speech responses to questions or responses to various surveys can be captured. Moreover, researchers can plan multiple surveys per day with specific notification times.

Both active and passive data are stored in *CSV* files. All data records or entries include a timestamp and unique hashed identifier corresponding to a patient.

Type of passive data	Description
Accelerometer	Accelerometer data from 3 axes x, y, z
GPS	Longitude, latitude and accuracy
Power state	Status of charging, battery level, lock and screen status
Gyro (iOS only)	Rotation or twist on a phone through the gyroscope
Reachability (iOS only)	Connection through Wi-Fi, cellular or unreachable internet
Magnetometer (iOS only)	Uncalibrated magnetic field of the phone in each direction
Device motion (iOS only)	Motion of the participants' phone taking into account gravity and exerting forces. Roll, pitch, yaw and rotation rate in three axes
Bluetooth (Android only)	Hashed mac addresses of nearby bluetooth devices
Wi-Fi (Android only)	Hashed mac addresses of nearby Wi-Fi routers

Table 3.1: Beiwe passive data

Type of active data	Description
Audio sample	Speech recording
Survey	Numerical slider response
Survey	Response using list with options
Survey	Checkbox response
Survey	Free response

Table 3.2: Beiwe active data

3.2.3 Data Privacy and Security

Privacy and security are significant issues, and Beiwe tries to handle these problematic points according to the latest policies. The data are stored on the phones temporarily, cached, and uploaded to the backend server every time the phone reaches a Wi-Fi connection. The upload option through a cellular data can be turned on, but it is not recommended.

All identifiers are protected by hashing, and all data are encrypted at all times. Moreover, the platform creates a public key used to encrypt all the data and a private key used to decrypt data on the server. Thus, the data cannot be accessed even if the device is stolen as the Beiwe application cannot read its data. Furthermore, a new unique authentication key is needed to download the data, which administrators of a study can create. [TKLO16]

3.2.4 Deployment

Deployment of the Beiwe platform is a specific issue, and one can find multiple obstacles regarding the Amazon Web Cloud Services. Thus, I have created a Beiwe deployment manual available as Appendix C, which also considers price and scalability.

3.2.5 Usage

Beiwe platform can be considered as a beaten track. Thus, this platform's use in studies focusing on digital phenotyping is quite common. For example, we can find the platform used to collect data to identify and track speech decline in amyotrophic lateral sclerosis [CGP+19]. Another use regarding the preceding disease was in quantifying its progression [BPC+19]. Moreover, Beiwe was used to quantify mobility and quality of life within patients with spine disease [CBOS19].

Further, it was used to predict loneliness and companionship based on geosocial features [WBC+21]. Another promising pilot study was focused on enhancing cancer care [WRS+18]. The application's use is also common regarding mental illnesses, and a study focusing on schizophrenia is not an exception [TSB+18].

3.3 MindLAMP

3.3.1 Description

The second data collection platform that is fairly new is MindLAMP. It was introduced in 2019 by Harvard Medical School in Boston. However, this platform is not a project of one specific department or similar entity. The application's development lies on multiple institutions' shoulders and can be seen as a multi-institutional research project with a Boston home. Though the basic functioning and architecture are similar to Beiwe, it might be considered a distinctive digital phenotyping approach. The idea behind the MindLAMP started with the need to provide a dashboard for researchers, a dynamic application for patients and a backend system to store data and concurrently to offer an immediate advantage for patients such as visualisation or insight into their data. There was also a need for flexibility. In other words, having a mobile application that can be customised, open-source, transparent, and can deliver some added value for patients. [TWB+19]

When developing the system, the following initial aims were taken into account. Firstly, there was a need to create a translational application, which would connect clinics, patients and researchers, based on clinical relevance. It was made together with psychiatrists, psychologists, doctors and patients. Secondly, the aim was to provide transparent code with as much customisation for users as possible. In other words, it should be an open and reusable application for various sets of medical problems. Thirdly, it was necessary to introduce a new type of data that could give researchers unique insights into mental illnesses. [TWB+19]

One could say that all of the aims were completed. Currently, it is a partly functional platform, which is still in development, with the backend solution, application available on iOS and Android mobile operating systems, dashboard and patients' portal. However, some parts are not fully completed, and updates are deployed weekly.

Interestingly, the man behind the project, *John Touros, MD MBI,* is the same man who helped develop the Beiwe application in its early and middle years and was behind the whole idea. However, after the call with its team, I discovered that creating of the MindLAMP platform started because of the different views regarding treating mental illnesses that Beiwe represents. The biggest problem was that the Beiwe application could not offer immediate intervention or visualisations for patients, and it is instead a data collection system only.

3.3.2 Data Collection

MindLAMP collects raw data to create a homogenous data set across Android and iOS mobile operating systems. As shown in Table 3.3, the data the application collects is unified across the two mobile platforms. In other words, it does not matter which mobile operating system patients use, as the data types produced by MindLAMP are similar. Only information from calls and

Type of passive data	**Description**
Accelerometer	Accelerometer data from 3 axes x, y, z
GPS	Longitude, latitude and accuracy
Power state	Status of charging, battery level, lock and screen status
Gyro	Rotation or twist on a phone through the gyroscope
Magnetometer	Uncalibrated magnetic field of the phone in each direction
Device motion	Motion of the participants' phone taking into account gravity and exerting forces. Roll, pitch, yaw and rotation rate in three axes
Wi-Fi	Hashed mac addresses of nearby Wi-Fi routers
Calls (iOS only)	Log of incoming and outcoming calls with their length
Distance moved (iOS only)	Number of steps

Table 3.3: MindLAMP passive data

Type of active data	Description
Survey	Response using list with options
Survey	Free response
Cats and dogs	Game for tracking speed of responses
Spatial span	Game for tracking speed of responses
Jewels trials	Game for tracking speed of responses

Table 3.4: MindLAMP active data

the number of steps is mobile platform-specific. A considerable disadvantage is the inability to control sensors' sampling frequency. Therefore each device produces a different amount of data per sensor. The sampling is currently set up for "as much as it can", but this issue will be fixed soon. However, turning off sensors for a study is achievable.

As shown in Table 3.4, the active data collected by MindLAMP consist of two formats of surveys and data from playing games. Consequently, the MindLAMP platform introduces metadata, a new type of data that the application can collect. These data could provide insights into the specific behaviour of an individual. For example, it can be an amount of time to fill out a survey or react to games available in the MindLAMP application. However, this option is still in development and will be deployed soon. Further, there is a possibility for a patient to see their data visualised.

All data entries are stored in *JSON* format and are downloaded through a script or research dashboard. All data rows contain a hashed identifier and timestamp to identify a device.

3.3.3 Data Privacy and Security

MindLAMP is a new platform that significantly considers data privacy and security. During the development, three core themes, trust, control and community, were considered. It means that the patients want to have control over their data. Therefore, all of their data can be erased from the application and servers if wanted. Moreover, all data are encrypted, and identifiers are hashed to protect users' privacy. [TWB+19]

3.3.4 Usage

Because the MindLAMP platform is a new project and is still in active development, this platform's utilisation is relatively rare. Nonetheless, we can still find pilot studies where it was used. As proof, the platform was used to find the relationship between digital media use and mental health [Nis20]. Though it is not currently heavily used, it appears as an up-and-coming solution for the future where only MindLAMP might be used instead of multiple apps focusing on digital health technology [TV20]. Besides, the platform could deliver just-in-time and longitudinal intervention for patients with clinical problems [VHT19].

3.4 Other Platforms

Other platforms that can be considered for data collection might not be that complex as Beiwe or MindLAMP. For example, a PRIORI app that focuses just on speech analysis of sound waves was used to predict mood states [MGMP17]. Moreover, architecture DeepMood was used for mood detection using mobile phone typing [CZZ+17]. However, there was a need to give the participants adjusted mobile phones with the corresponding applications in both cases. A similar solution to Beiwe and MindLAMP was an application called Monarca which is, however, out of date and probably monetised [FMH+11].

Chapter 4

Comparison of Platforms

4.1 Introduction

This section compares two already mentioned platforms, Beiwe and MindLAMP, currently used among researchers and available to collect behavioural data from mobile phones. I used the following criteria to find the suitable and the most solid solution:

1. Deployment of both platforms
2. Testing of the systems
3. Pilot study across six individuals
4. Collection of data from both platforms for two weeks
5. Creation of detailed graphs
6. Big picture view and final detailed comparison

The summary of components that were compared can be found in Table 4.1. I will support my conclusions, row by row, in the following sections. These conclusions come from a series of testing of Beiwe and MindLAMP platforms on 3 iOS devices and 3 Android devices. Eventually, two week period was used for simultaneous data collection after multiple issues with servers and mobile phones were solved.

4.2 Research

4.2.1 Price, Hosting and Deployment

With the Beiwe platform, it is necessary to rely on Amazon Web Services infrastructure. However, the server solution from Amazon can be costly. According to my calculation, sufficient infrastructure for 20 patients can cost up to 200 dollars per month. Moreover, no manual is offered, and thus the users of Beiwe have to figure out the proper server setup on their own.

Table 4.1: Comparison of platforms

		Beiwe	MindLAMP
RESEARCH	Price	100$/20 users/1 month	Free of charge
	Hosting	AWS	Own server
	Deployment/management	AWS, simple	Own server, more complex
	Pipeline/preprocessing	Not working	Missing
	Data obtainment	Simple, intuitive	More complex
	Dashboard	Fully operational	Working with limitations
	Adjustability	OK	Missing features
	Questionnaires	Simple creation	More difficult creation
UI & UX	Onboarding	Intuitive/supervision	Intuitive/supervision
	Battery consumption	5% per day	5-25% per day
	Ease of use	Interaction not necessary	Big difficulties
	App runtime errors	No occurrences	Runtime errors, logging out
	Background activity	Not necessary	Essential
	Questionnaires	Simple interaction	More complex interaction
DATA	Data types	Less data types	More data types
	Data quality	Small difficulties	Big difficulties
	Sampling frequency	Adjustable, unified	Unadjustable, varied
	Data collection outages	OK	Occasional
	Means of sending	Wi-Fi/cellular	Wi-Fi/cellular
	Caching	Yes	Limited

Legend - Quality: Good | Small difficulties | Big difficulties | Bad | Missing

Assuming sufficient infrastructure is available, MindLAMP can be deployed without additional costs. On the other hand, Amazon Web Services are reliable and have working notifications about outages. As for MindLAMP, the server's possible problems must be solved by the infrastructure administrator, and notifications need to be implemented separately. Moreover, Beiwe can be more simply deployed than MindLAMP if deployment instructions from my manual in Appendix C are used.

4.2.2 Preprocessing

Beiwe platform offers a pipeline for downloading and also preprocessing data. However, the pipeline does not work currently. MindLAMP lacks this feature too. Novertheless, both problems should be fixed soon based on communication with both platforms' developers.

4.2.3 Data Obtainment

For Beiwe, data can be obtained using a few clicks in the dashboard, while for MindLAMP, a relatively complex script must be executed to download data from the database.

4.2.4 Dashboard

Dashboard of Beiwe works well and is well-arranged, although it is not graphically captivating. MindLAMP's dashboard has more options and is visually attractive. However, it contains bugs.

4.2.5 Adjustability

Studies based on the Beiwe platform are entirely adjustable as the sampling frequency can be chosen and edited. Moreover, the option to upload data via cellular or Wi-Fi can be selected. These options are not yet possible with the MindLAMP platform.

4.2.6 Questionnaires

While MindLAMP offers a broader range of options, questionnaires' creation is more complex than with Beiwe, where it is possible to create a survey using a few clicks.

4.3 User Interface and User Experience

4.3.1 Onboarding

There is a simple onboarding on both platforms for a patient. With corresponding details, the patient can log in for a study. However, patients with more severe problems might require supervision or phone help.

4.3.2 Battery Consumption

Beiwe mobile application uses significantly less battery than the MindLAMP application. However, the exact percentage depends on the specific device and sensors collected.

When collecting similar data on the same devices, the battery consumption on MindLAMP can be up to five times bigger. Typically, the Beiwe application consumes 5% of battery capacity per day, and MindLAMP can easily consume up to 25%.

4.3.3 Ease of Use

Beiwe application is reliable and does not need to be checked after initial setup. Moreover, the application does not have to be controlled or operated during studies if surveys are not used. In contrast, the MindLAMP application is not as easy to get yourself oriented in because users can navigate the application, play games, or see their data visualisations. Further, frequent errors cause a user to log out, resulting in need to log in again.

4.3.4 Application Errors

There were no runtime errors or crashes on the Beiwe mobile application. However, the MindLAMP application's frequent errors on iOS caused a user to log out involuntarily. Moreover, the application had to be uninstalled and installed again because of the errors. Furthermore, when the error happens, the MindLAMP immediately stops sending data.

4.3.5 Background Activity

While the Beiwe application does not have to be kept in the active applications to upload or collect data, the MindLAMP application must be running in the background to collect and upload data. This problem is especially significant on iOS devices. Moreover, some Android mobile manufacturers use automatic cancellation of active applications, which is not wanted when MindLAMP runs.

4.3.6 Questionnaires

Due to the more complex nature of the MindLAMP app, filling out questionnaires might be more difficult for users with mental disorders. When compared with the Beiwe application, there is no other possibility than filling out the survey on the application dashboard.

Passive Data	Beiwe	MindLAMP
Accelerometer	X	X
GPS	X	X
Power State	X	X
Gyro	iOS only	X
Reachability	iOS only	-
Magnetometer	iOS only	X
Devicemotion	iOS only	X
ios_log	iOS only	-
Bluetooth	Android only	-
Wi-Fi	Android only	X
Android_log	Android only	-
Calls	-	iOS only
Distance moved	-	iOS only

Table 4.2: Passive data comparison

4.4 Data

4.4.1 Types and Quality

Beiwe platform is able to collect fewer types of data than MindLAMP. Moreover, as shown in Table 4.2, MindLAMP is more robust as it gathers similar data simultaneously from Android and iOS mobile devices. Nevertheless, the data quality is lower for MindLAMP due to the already mentioned application errors and a need to run in the background. Moreover, as shown in Table 4.3, Beiwe can collect voice surveys but does not offer games for collecting metadata.

Active Data	Beiwe	MindLAMP
Voice survey	X	-
Multiple choice survey	X	X
Free text survey	X	X
Slider survey	X	-
Checkbox survey	X	-
Spatial span game	-	X
Cats and Dogs game	-	X
Jewels Trials A & B game	-	X

Table 4.3: Active data comparison

4.4.2 Sampling Frequency and Collection Outages

A considerable advantage of Beiwe is the possibility to edit the sampling frequency of sensors gathering data. On the contrary, it is impossible to adjust data sampling frequency on MindLAMP, and the native one varies across devices. Furthermore, data sets from MindLAMP contain more missing values than from Beiwe due to the application runtime errors and crashing.

4.4.3 Means of Sending and Caching

Both platforms offer the ability to upload data via Wi-Fi or cellular. This option can be freely chosen. Nevertheless, upload of data via cellular does not work currently on MindLAMP. Moreover, when the cellular or the internet connection is unavailable, Beiwe uses caching and sends data when the internet connection is available again. This option is supported by MindLAMP too, but just with limited time windows which are often not long enough and result in data corruption.

4.5 Data Types in Detail

4.5.1 Introduction

As shown in the Table 4.4, this section elaborates on data quality from specific sensors and describes the data quality reasoning. Moreover, if making sense, visualisations of data are added. During the pilot study and testing, the following devices were used corresponding to specific users:

- iOS: *iPhone - 11 Pro (User 1)*, *XR (User 2)* and *8+ (User 3)*
- Android: *Xiaomi Mi Max 2 (User 4)*, *Samsung Galaxy S8+ (User 5)* and *Xiaomi Redmi 4X (User 6)*

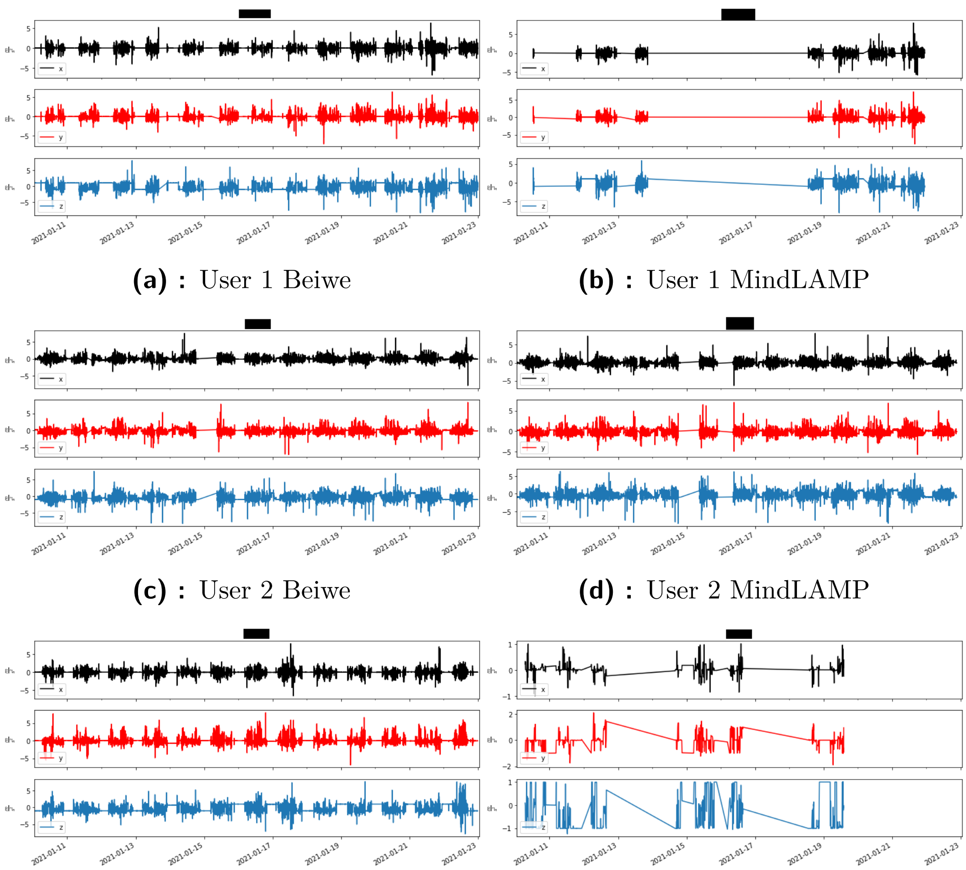
4.5.2 Accelerometer

- `Beiwe`
  - *iOS*
    As shown in Figure 4.1, there are high-quality data sets with both excellent and consistent sampling frequency. Data collection outages are infrequent, and values are comparable across all iPhone models used in the data collection process.
  - *Android*
    There are significant differences in data quality across devices, as reflected in Figure 4.2. For example, there is an overly high sampling frequency on *Xiaomi Mi Max 2*, an overly low sampling frequency on *Xiaomi Redmi 4X*, and there are only a few isolated data points on *Samsung S8+* with large segments completely missing.

	Beiwe		MindLAMP	
	iOS	Android	iOS	Android
Accelerometer	Good	Big difficulties	Big difficulties	Small difficulties
GPS	Good	Good	Big difficulties	Big difficulties
Bluetooth	Missing	Good	Bad	Missing
Wi-Fi	Missing	Good	Big difficulties	Small difficulties
Devicemotion	Small difficulties	Missing	Small difficulties	Small difficulties
Gyroscope	Good	Missing	Small difficulties	Small difficulties
Magnetometer	Bad	Missing	Good	Good
Reachability	Big difficulties	Missing	Missing	Missing
Power State	Good	Good	Big difficulties	Big difficulties
Calls	Missing	Missing	Good	Missing

Quality: Good Small difficulties Big difficulties Bad Missing

Table 4.4: Data quality comparison



(a) : User 1 Beiwe

(b) : User 1 MindLAMP

(c) : User 2 Beiwe

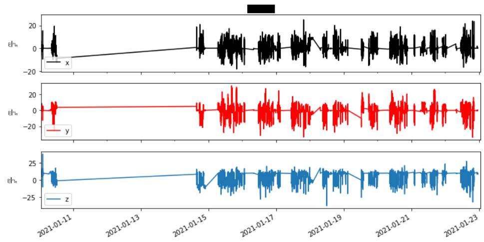
(d) : User 2 MindLAMP

(e) : User 3 Beiwe

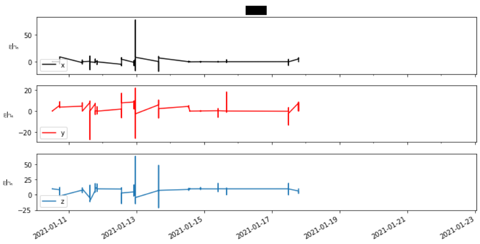
(f) : User 3 MindLAMP

Figure 4.1: Accelerometer iPhone comparison

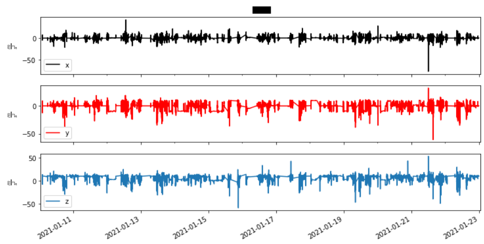
(a) : User 4 Beiwe - Too much of data points



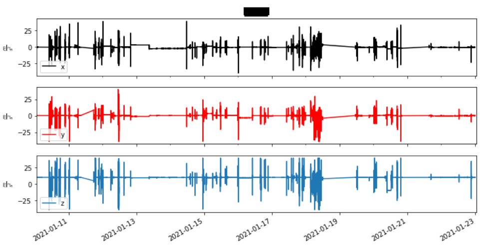
(b) : User 4 MindLAMP



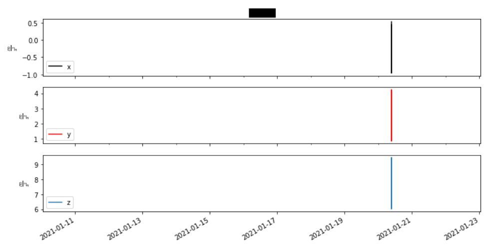
(c) : User 5 Beiwe



(d) : User 5 MindLAMP



(e) : User 6 Beiwe



(f) : User 6 MindLAMP

Figure 4.2: Accelerometer Android comparison

- **MindLAMP**
  - *iOS*
    As shown in Figure 4.1, there are extreme data quality differences across devices. As for *iPhone XR*, the data are comparable to the data collected using Beiwe. On *iPhone 8+*, the sampling frequency is shallow, and on *iPhone 11 Pro*, there are significant time segments without any observations.
  - *Android*
    Figure 4.2 shows that reliable data were collected using *Xiaomi Mi Max 2* and *Samsung Galaxy S8+* with an excellent sampling frequency. However, there was an extensive data collection outage for the first of the named devices. As for *Xiaomi Redmi 4X*, data are absent except for a few data points obtained within a single hour from the two weeks of the data collection process.

4.5.3 GPS

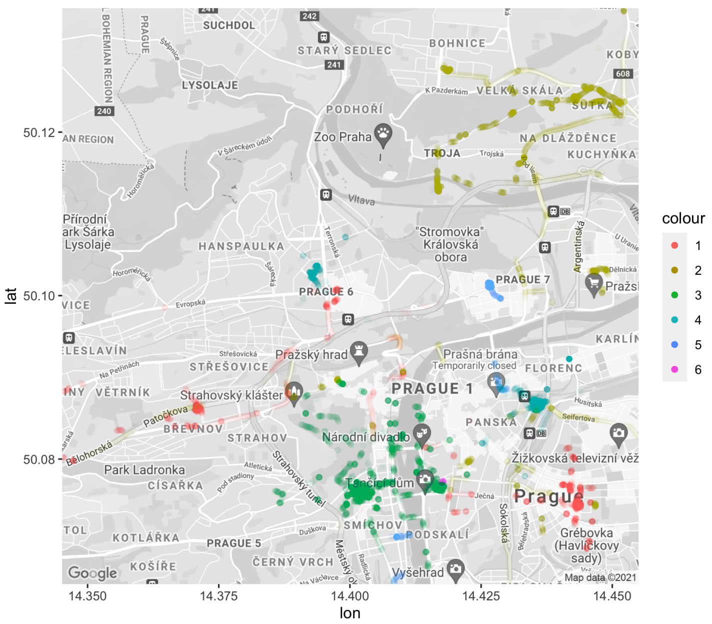
- `Beiwe`
  - *iOS & Android*
    GPS sensor produces very high-quality data, and these data are collected in regular intervals that can be adjusted. There is a sufficient number of data points for each device without outages, as reflected in Figure 4.3 and Figure 4.4. Smoother paths could be reached by increasing sampling frequency.

- `MindLAMP`
  - *iOS & Android*
    The data are of decent quality, but there is no consistency across devices from the sampling frequency perspective. Furthermore, the fact that the data are not well cached when the user is not connected to the internet represents a massive obstacle for GPS data simply due to their nature and information value. As proof, Figure 4.3 and Figure 4.4 show that there are more data points with MindLAMP because of the sampling frequency, but some paths are entirely missing.

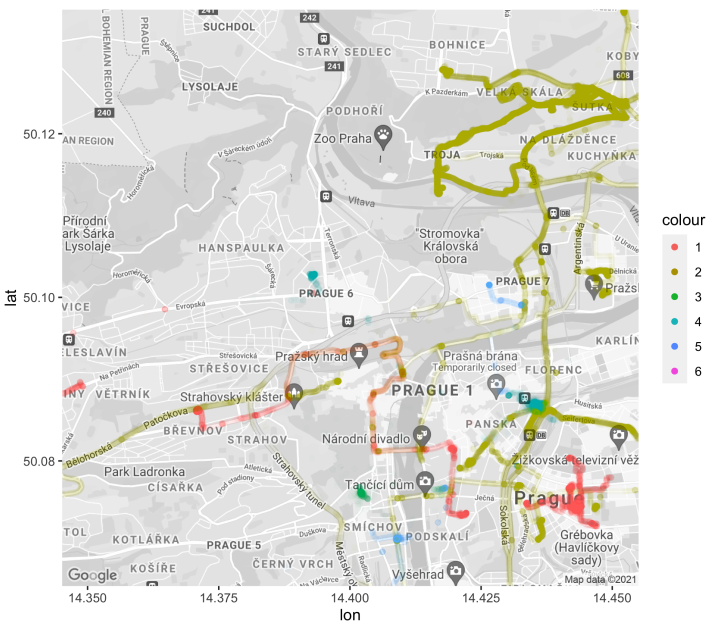
4.5.4 Bluetooth

- `Beiwe`
  - *iOS* - Missing
  - *Android*
    Data are of outstanding quality for all three Android devices included in the data collection process. In Figure 4.5 we can see number of unique devices per period of the pilot study.

- `MindLAMP`
  - *iOS*
    While the documentation of MindLAMP does not mention Bluetooth data collection on iOS, several observations were collected. However, the data frame is insufficient for work as there are insufficient data points for a proper analysis.
  - *Android* - Missing

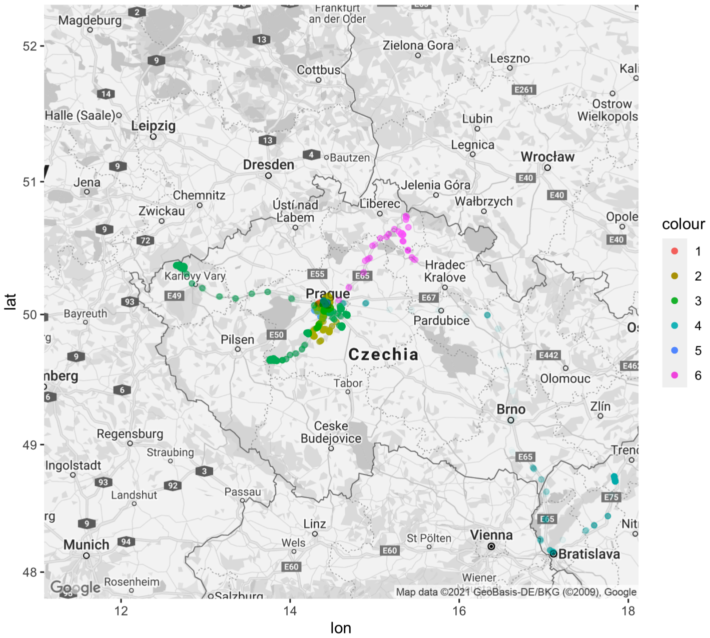


(a) : All users Prague Beiwe

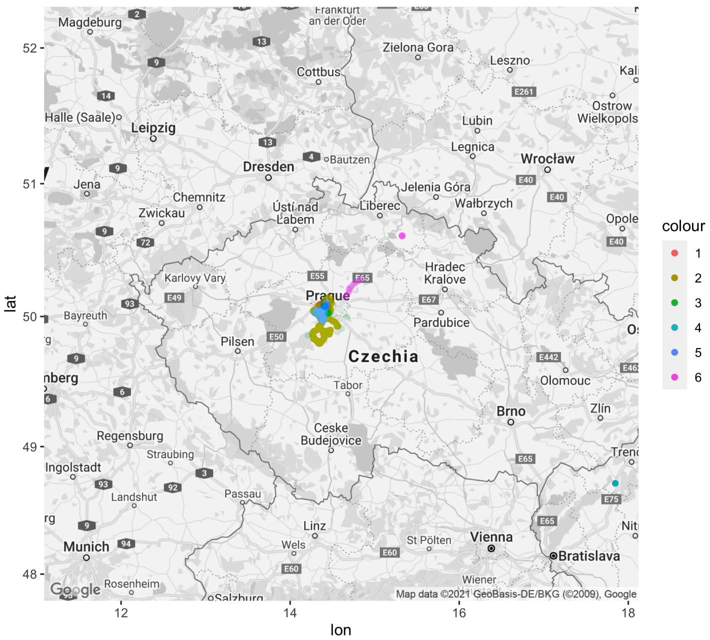


(b) : All users Prague MindLAMP

Figure 4.3: GPS data comparison: "Prague, Czechia" Map, Google Maps.



(a) : All users Beiwe



(b) : All users MindLAMP

Figure 4.4: GPS data comparison: "Czechia" Map, Google Maps.

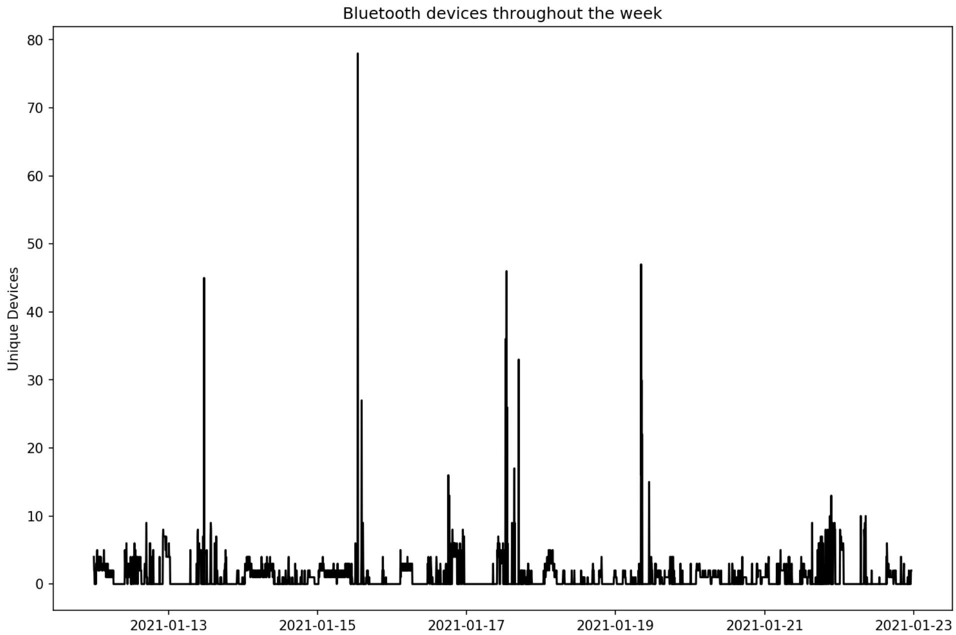


Figure 4.5: Bluetooth data example User 5

4.5.5 Wi-Fi

- `Beiwe`
  - *iOS* - Missing
  - *Android*
    Wi-Fi data are sufficient for two of the three Android devices included in the data collection process.
- `MindLAMP`
  - *iOS*
    Low-quality data cannot be used due to a low sampling frequency.
  - *Android*
    There were stable data on *Samsung S8+*, a disproportionate number of data points on *Xiaomi Mi Max 2* and insufficient data on *Xiaomi Redmi 4X*.

4.5.6 Devicemotion

- `Beiwe`
  - *iOS*
    The data contains magnetometer, user acceleration, rotation, device motion, and gravity. There is a sufficient amount of data where a user activity can be very well estimated. Though, sometimes some columns contain *NULL* values.

  - *Android* - Missing
- `MindLAMP`
  - *iOS*
    The collected data include acceleration, altitude, rotation, gravity, and magnetometer, where altitude data correspond to the data obtained using Beiwe. However, *iPhone 8+* did not collect acceleration and altitude data during the data collection period. *iPhone X* and *iPhone 11 Pro* did not collect data from the magnetometer. These pieces of data are, however, collected using the magnetometer sensor.
  - *Android*
    There are only rotation and magnetometer data collected.

4.5.7 Gyroscope

- `Beiwe`
  - *iOS*
    Good data quality. However, the sampling frequency differs across different models of iPhones.
  - *Android*
    While the Beiwe documentation does not mention gyroscope data collection on Android devices, data were obtained from two out of the three Android devices used in the data collection process. However, the data points were highly infrequent.
- `MindLAMP`
  - *iOS*
    There are very significant differences in sampling frequency.
  - *Android*
    There is a solid data quality on *Xiaomi Mi Max 2* and *Samsung Galaxy S8+* and good sampling frequency. However, on *Xiaomi Redmi 4X*, data are absent.

4.5.8 Magnetomer

- `Beiwe`
  - *iOS* The data contain only NULL data points. Thus, the data are not usable.
  - *Android* - Missing

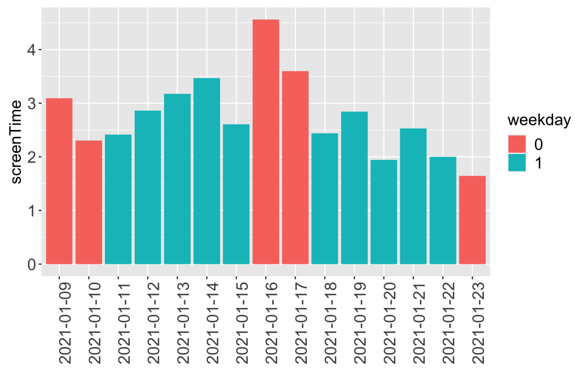
- `MindLAMP`
  - *iOS*
    There is a sufficient amount of data, easily read and visualised. The sampling frequency is relatively low for *iPhone 8+*. On the other hand, the newer models are sending data almost constantly.
  - *Android*
    The magnetometer itself does not contain any data for Android. However, it is possible to obtain them from the device motion sensor.

4.5.9 Reachability

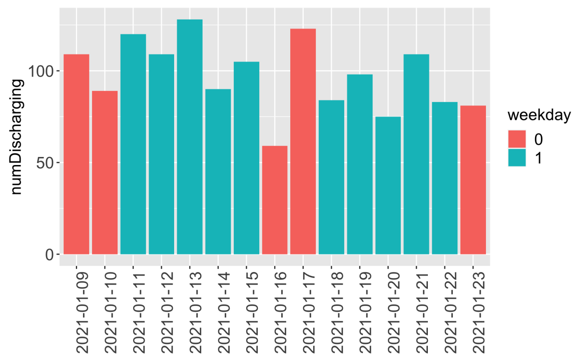
- `Beiwe`
  - *iOS*
    There is a sufficient number of observations in the data. However, they do not seem to have any information value. At various times, information pieces about connection to Wi-Fi, cellular, or no network are collected. This collection, nevertheless, does not correspond to changes in connection type since the data points are usually seconds apart, and the connection type would thus change suspiciously often.
  - *Android* - Missing
- `MindLAMP`
  - *iOS* - Missing
  - *Android* - Missing

4.5.10 Power State

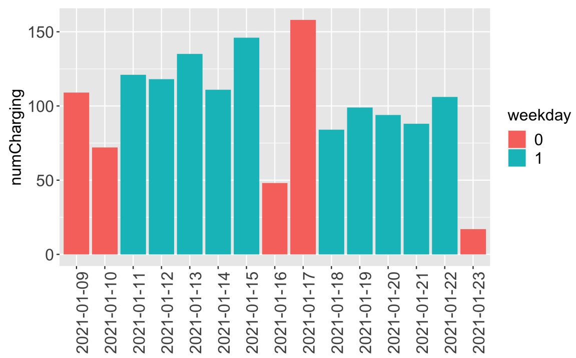
- `Beiwe`
  - *iOS*
    The data are sufficient and produce a reliable data set from the two mobile software platforms. As shown in Figure 4.6, screen time and total level of battery charged or discharged can be found out from these data points.
- `MindLAMP`
  - *iOS & Android*
    Typically, dozens of *Screen OFF* observations in a row should not be possible since the display should first be turned on before it can be turned off. This inconsistency is not specific to a particular device. Therefore, it is likely a problem with mindLAMP itself.



(a) : Screen time in hours



(b) : Battery percentage points discharged



(c) : Battery percentage points charged

Figure 4.6: Power state data example User 2

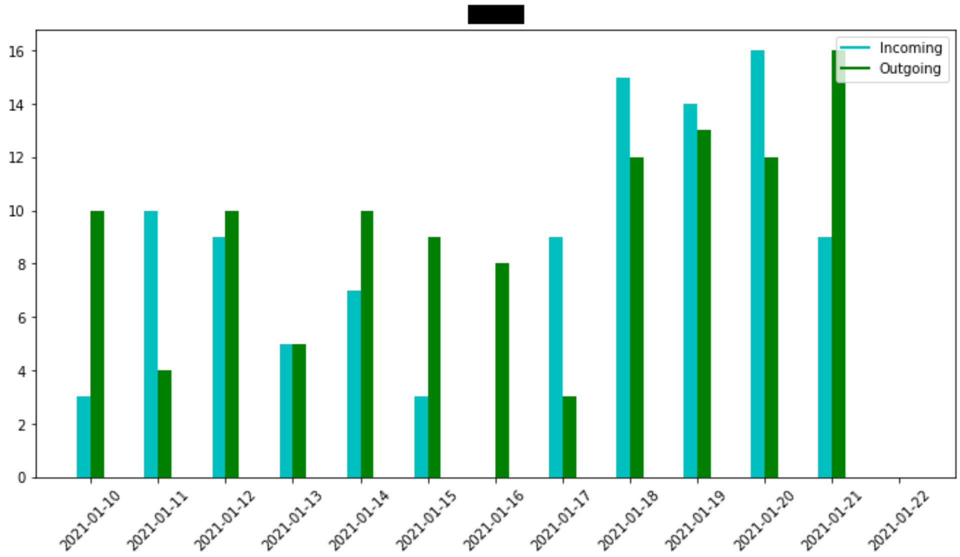


Figure 4.7: Number of calls data example User 2

4.5.11 Calls

- `Beiwe`
  - *iOS & Android* - Missing
- `MindLAMP`
  - *iOS*
    The data are of exceptionally high quality, with a minor exception of data from *iPhone 8+*, where multiple call observation duplicates are created.
  - *Android* - Missing

4.6 Summary

4.6.1 Data Set

Beiwe produces a more stable and valid data set because it supports caching in a great deal compared to MindLAMP, and the frequency of each data type can be adjusted. In MindLAMP, there is no way to control the frequency of sampling, and each device collects data in different periods - as much as it can. However, Beiwe has problems with some types of Samsung phones. More importantly, MindLAMP does not work well on the iOS mobile platform as the app needs to be reinstalled when MindLAMP stops collecting data which happens nearly daily. Thus, there is a lot of missing data compared to the Beiwe data set.

4.6.2 Researcher Perspective

From a researcher and a maintainer perspective, there is less maintenance with the Beiwe platform as AWS sends notifications when the server has problems and can be considered, in general, more stable than MindLAMP's server solutions. Additionally, I have experienced multiple issues with collecting data on MindLAMP as it sometimes stopped collecting data on the iOS platform, and it needed to be reinstalled. Furthermore, the MindLAMP application has to run in the background continuously. Thus, there needs to be a researcher who controlls the data streams daily and, if necessary, notifies the users with broken data streams to start the application.

4.6.3 Patient Perspective

Beiwe application's user interface is very simple as the patient can only fill in ongoing surveys, and it is not that visually pleasant. On the other hand, the user interface of MindLAMP is visually delightful but might be too complex and too detailed for a patient with some mental health disorder as there exists a possibility to play games or look at data visualisations. Nevertheless, this might be an appropriate question for a psychologist or psychiatrist with the relevant education and background.

4.6.4 Current and Further Development

Beiwe is still functional but an older platform for digital phenotyping. Nonetheless, it lacks vision, agile approach for bugs and problems. Additionally, it relies on Amazon Web Services, which are stable but expensive. However, MindLAMP is a new app with flaws such as malfunctioned caching, nonexistent data frequency setup and three times more extensive battery consumption. Nevertheless, it is updated constantly in all attributes such as the backend, the app itself, and the documentation. Thus, currently, Beiwe is a more convenient option, but MindLAMP might be soon the way to go.

Chapter 5

Pilot Study

5.1 Introduction

This chapter will describe a pilot study that was started after the preceding platform comparison pilot study. After a thorough decision, the research team concluded that the best platform to collect behavioural data from smartphones is currently Beiwe. Though it offers less variety and smaller number of data types, it brings stability, foresight and a couple of years of experience. However, this situation might change in a few months when MindLAMP is a more stable platform with all the problems fixed.

5.2 Actigraphy

In order to obtain as much data as possible about an individual, it was decided to collect actigraphy data. *Actigraphy* as a phenomenon, in general, refers to a method using wristwatch-like devices to monitor and collect data generated by a person's movement [SA02].

The actigraphy data are essentially a time series with a fixed sample rate where each sample corresponds to a particular value. The data collection in our study was achieved by a special wristband called MindG with an accelerometer included and provided, as already mentioned, by MindPax Ltd. This device collects signals from three axes *(x, y, z)*, and then the magnitude of the acceleration vector is computed as a squared root of the sum of the signals from the x, y and z axes squared. The acceleration vector with the highest magnitude is chosen every 30 seconds, and then the data stream is sent to the application when synchronising. [AVN26]

Actigraphy can be considered as a profound method that is relatively old compared to the phenomenon of digital phenotyping, which is discussed in this thesis. However, I believe that actigraphy can be considered a part of digital phenotyping. Moreover, actigraphy is used vastly across the professional community, determining patterns in mental or physical health disorders. As proof, actigraphy can be used with great success in assessing the sleep or circadian rhythms of people [SA02]. These features might help assess bipolar patients outside of acute manic or depressed episodes [JHE05]. Moreover, actigraphy can detect behavioural patterns in schizophrenia [BHO+10]. Nevertheless, actigraphy plays a vital role as a good source of information about people with depression [BMS+13].

To sum up, accelerometer data from the wristband could provide a better understanding of a person's activity than just accelerometer data from the Beiwe platform.

5.3 Study Specification

Twenty participants with both Android and iOS mobile systems were asked to participate in the study for two months in order to be able to obtain a reasonably sized data set. Participants were asked to install the following applications to their smartphone:

- Beiwe
- Mindpax.me

Moreover, they were requested to fill in everyday surveys in the Mindpax.me mobile application. Included questions and possible answers can be seen in Table 5.1. As shown in Table 5.2, another survey was needed to be filled out weekly. Moreover, Table 5.3 shows possible answers to every statement.

Besides, it was required to wear a wristband that collects actigraphy data described in Section 5.2.

Question	**Type of answer**
When did you go to sleep last night?	Precise time
When did you wake up today?	Precise time
Mood	Scale from $[-1, 1]$ where -1 is the saddest and 1 is the happiest
Inner feeling	Scale $[-1, 1]$ where -1 is the most anxious and 1 is the most relaxed
Energy	Scale from $[-1, 1]$ where -1 is the least energetic and 1 is the most energetic
What was your last day routine?	Options: work from home, work from office, combination of both, free day
How much time did you spend exercising yesterday?	Precise time

Table 5.1: Everyday survey

Statement
I do not enjoy anything and nothing pleases me.
I have no energy.
I feel gloomy and pessimistic about the future.
I feel unusually great, optimistic.
I have excess energy.
My thinking is very fast, others cannot keep up with me.
I need to sleep less than usual.
I feel restless, tense.
I cannot focus.

Table 5.2: Weekly survey

Answer
I do not agree.
More likely I do not agree.
I probably agree.
I agree
I completely agree

Table 5.3: Weekly survey options

5.4 Acceptance Process

It was required by participants to go through and sign the following documents:

- Informed consent
- Consent to the processing of personal data
- Entry questionnaire

The entry questionnaire was in the form of Mini-international neuropsychiatric interview. This interview is a short clinical review of psychiatric patients, which allows to find out in a couple of minutes the patient's mental state according to *DCM-IV* [LSW^{+}97]. This interview provided necessary data on the participant's mental state and data set validity.

When the participant consented to the previous documents and was classified as healthy or did not suffer from depression or mania episode, they were included in the study. Simultaneously the participants were immediately assigned unique identifiers to protect their privacy. Thus since this time, they were referred by these unique identifiers.

In the next phase, each participant was assigned a researcher who helped with the installation of mobile applications, delivered the wristband, and guided the participant with the initial process.

Data collected daily	Origin
Accelerometer	Beiwe mobile application
GPS	Beiwe mobile application
Power state	Beiwe mobile application
Gyro (iOS only)	Beiwe mobile application
Reachability (iOS only)	Beiwe mobile application
Magnetometer (iOS only)	Beiwe mobile application
Device motion (iOS only)	Beiwe mobile application
Bluetooth (Android only)	Beiwe mobile application
Wi-Fi (Android only)	Beiwe mobile application
Actigraphy data	Wristband MindG
Time of falling asleep	Mindpax.me mobile application
Time of waking up	Mindpax.me mobile application
Energy level	Mindpax.me mobile application
Mood level	Mindpax.me mobile application
Anxiety level	Mindpax.me mobile application
Time of sport activity	Mindpax.me mobile application
Type of work	Mindpax.me mobile application

Table 5.4: Pilot study: data collected daily

5.5 Data Validation

As shown in Table 5.4, all the data that the Beiwe platform could obtain were collected. Moreover, actigraphy data from the wristband and answers to surveys were gathered. Above that, as can be seen in Table 5.2, collected answers to these statements were collected. However, not always were the participant filling out the surveys and sometimes there were problems with data streams in the mobile applications. Thus, assigned researchers took care of the data quality daily and, if necessary, contacted the participants. For Beiwe data validation, I created a script that visualised accelerometer and GPS data from the previous day and thus, possible outages and missing values were easy to detect.

Chapter 6

Data Preprocessing and Processing

6.1 Introduction

As my goal was to find relationships between GPS, power state, actigraphy data and mood, I will describe, in detail, the nature of these data and outline corresponding preprocessing needed and subsequently processing and features calculated using these data types. The data comes from 20 patients from the pilot study mentioned in the preceding chapter. The data were collected from a period from 15.3.2021 until 11.4.2021. Thus, the starting data set contains 28 days of raw data per patient. The total number of days collected was 488 days due to the late study start within some subjects. However, not all data are complete or valid, which I will outline in this chapter. Also, during the data analysis, it has to be taken into account that the data set is relatively small. Moreover, I will refer to the study participants as patients with a unique identifier *BA001* - *BA020*.

6.2 GPS Data

6.2.1 Characteristics

The first goal was to preprocess data from the GPS sensor. These data were downloaded from our Beiwe database, and they come in the form of CSV files, which are created hourly. These CSV files were concatenated and consequently preprocessed as one data stream per patient. As shown in Table 6.1, each data entry before preprocessing consists of longitude, latitude, altitude, accuracy, UTC time and timestamp. The GPS sensor works in a way that is either turned on or turned off. So when the sensor is on, it collects as much as possible data, which results in varying frequency across the mobile platforms and mobile manufacturers.

The setup for our pilot study was having a GPS sensor on for 60 seconds and consequently having the GPS sensor off for 600 seconds. This setup resulted in ten minutes off intervals and one-minute interval filled with data. However, some patients reported in this one minute time window one or two values and the others had dozens of data entries. Thus, I extracted the median of longitude, latitude, altitude, and time for each one-minute interval per patient, as shown in Table 6.1.

Time-stamp	**UTC Time**	**Latitude**	**Longitu-de**	**Altitude**	**Accuracy**
Unix	Y-m-d H:M:OS	[°]	[°]	[m]	[m]

UTC Time median	**Latitude median**	**Longitude median**	**Altitude median**
Y-m-d H:M:OS	[°]	[°]	[m]

Table 6.1: GPS data format before (1) and after preprocessing (2)

6.2.2 Data Imputation and Filtration

The data were not complete and thus were sometimes absent. This issue sometimes happened because of the accidental permission rejection in the Beiwe mobile application or the phone running out of battery, and thus the application could not collect data or upload already collected data. As shown in Table 4, the percentage of missing values varied across the patients and was excessively high. Due to the size of the data set, I decided not to exclude corrupted patients immediately but to impute data first.

All empty time windows were imputed with the following procedure. Firstly, if the empty time window happened during the night, specifically between 9 pm and 9 am, and the last known value of longitude and latitude before the empty section and first known value of longitude and latitude after the empty part were in 200 metres distance, then the data were imputed by the mean value of time window endpoints demonstrated in equations below.

The same mean value imputation was used for time windows that were at most 150 minutes long during the day, and the distance of the endpoints differed maximally by 200 metres. As shown in Table 4, this procedure provided a relatively high correction to the patients with valid data. On the other hand, patients who had corrupted data, most likely because of the malfunctioned device, were not corrected and instead stayed within a high percentage of missing data. A similar imputation procedure was used in a study focusing on detecting bipolar depression from geographic location data [PTS+16].

$$meanLatitude = (laS + laE)/2 \tag{6.1}$$

$$meanLongitude = (loS + loE)/2 \tag{6.2}$$

$$meanAltitude = (alS + alE)/2 \tag{6.3}$$

Where:

laS = first known median latitude after the empty section
laE = last known median latitude before the empty section
loS = first known median longitude after the empty section
loE = last known median longitude before the empty section
alS = first known median altitude after the empty section
alE = last known median altitude before the empty section

Filtration of the data was also necessary. Thus, all the patients who had missing data percentage bigger than 40% after the imputation were excluded from the data set. This procedure resulted in excluding two patients who probably had malfunctioned devices. Moreover, one day of patient was excluded from the rest of the data sample if the missing percentage of the data per day was more prominent than 40% for a patient. This procedure resulted in data set containing 18 patients with 417 days of data after 61 days were excluded. The data set contains, in the end, a minimal amount of missing longitude and latitude data.

Patient	**% of missing data before imputation**	**% of missing data after imputation**
BA001	26%	6%
BA002	43%	17%
BA003	5%	2%
BA004	43%	23%
BA005	48%	25%
BA006	12%	3%
BA007	60%	50%
BA008	48%	32%
BA009	57%	37%
BA0010	53%	19%
BA0011	46%	24%
BA0012	53%	30%
BA0013	52%	31%
BA0014	18%	3%
BA0015	25%	23%
BA0016	24%	5%
BA0017	76%	73%
BA0018	50%	26%
BA0019	33%	9%
BA0020	30%	15%

Table 6.2: GPS data missing percentage before and after imputation

6.2.3 Features Extraction

To begin with, I would like to emphasise that all features were calculated for each patient separately. Thus, each patient corresponds to one data set and each row corresponds to one day.

Firstly, distances and time between each two data rows for each patient were computed. From this data, each row was classified as either stationary or dynamic. The criterion was the speed between points. The threshold was set to 2 kilometres per hour. In other words, if the movement speed of a subject was greater than 2 kilometres per hour between two points, the corresponding points were classified as dynamic and otherwise stationary. This classification of points was necessary for clustering. The data set of each patient was aggregated to days and contained the following features:

1. Total distance

   This feature was computed as a sum of distances between points in a day demonstrated below.

   $$\sum_{i=1}^{n-1} \Delta d_i \tag{6.4}$$

   where:

   Δd_i = change in the distance between two rows
   n = number of rows in one day

2. Location variance

   This feature was used in the study regarding correlation of depressive symptom severity in daily life. The logarithm was used to compensate for skewness in the distribution of location sample variance across participants. [SZK+15]

   $$\log_e(\sigma^2_{lat} + \sigma^2_{long}) \tag{6.5}$$

   where:

   $\log_e$ = natural logarithm
   σ^2_{lat} = sample variance of latitude data in stationary states
   σ^2_{long} = sample variance of longitude data in stationary states

3. Number of clusters

   To determine the number of clusters visited it was necessary to choose and run a clustering algorithm. I used hierarchical clustering with the complete linkage method [Jar20]. This algorithm aggregates similar objects into groups. In this case, it grouped spatial points into groups with the threshold set to 400 metres. Because of the complete linkage method, clusters with a distance of their furthest points smaller than 400 metres are always linked. Then the data were aggregated based on days. Moreover, an example of clustering can be found in Figure 6.1.

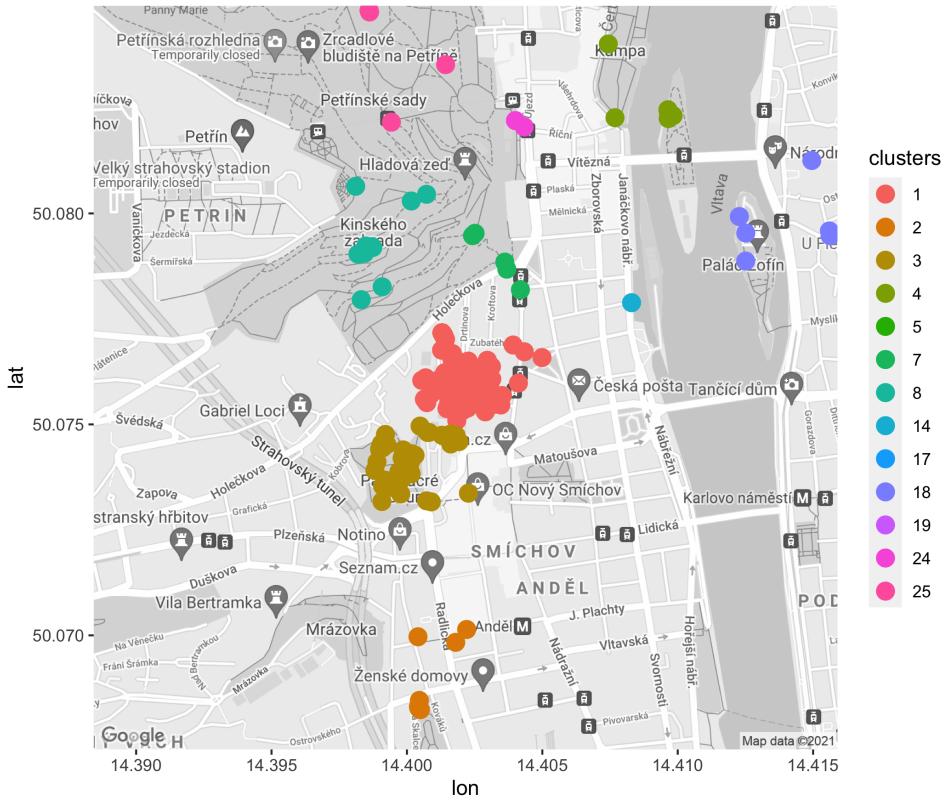


Figure 6.1: Participant's BA005 clusters: "Prague, Czechia" Map, Google Maps. Accessed 2021-4-28.

4. Time in each cluster

   Total time per specific cluster in one day was also computed. This computation resulted in multiple variables with a varying number of these variables across the patients.

5. Total time spend in clusters

   Total time spend in clusters was computed as a sum of time spent in each cluster per day. It can also be expressed as time spent in stationary states per day.

$$\sum_{i}^{n} c_i \tag{6.6}$$

where:

c_i = time spent in cluster i in a day
n = total number of clusters visited in a day

6. Entropy

In order to measure variability in the time that participants spent in different stationary location, entropy can be a quite useful feature. This feature is based on information-theoretical entropy [SEM91]. Moreover, precise definition of this variable, that will be described below, was already used in studies focusing on geographical data and mood or depression correlation [SZK+15, PTS+16, SLS+16, MQRP13].

$$-\sum_{i}^{n} p_i log_e p_i \tag{6.7}$$

where:

n = total number of clusters visited in a day
p_i = percentage of time the participant spent at the location cluster i

The bigger the entropy, the more uniform is the distribution of time spent in clusters. The smaller the entropy, the less uniform is the distribution of time spent in clusters.

7. Normalised entropy

Normalised entropy is essential as it does not relate to the total number of clusters. Moreover, normalised entropy ranges from 0 to 1 values. This fact typically means that normalised entropy of 0 means that the patient spent the whole day in one location. On the other hand, value 1 might reflect that the time between all clusters visited per day is uniformly distributed. [SZK+15, PTS+16, SLS+16, MQRP13]

$$(-\sum_{i}^{n} p_i log_e p_i)/log_e n \tag{6.8}$$

where:

n = total number of clusters visited in a day
p_i = percentage of time the participant spent at the location cluster i

6.3 Power State Data

6.3.1 Characteristics

These data came in the format of CSV files from our Beiwe database. Corresponding CSV files were concatenated and then preprocessed as one data stream per patient. However, there was a need to distinguish between iOS users and Android users because these two mobile platforms detect different screen events. This issue was solved by logs that Beiwe sends together with data. As can be seen in Table 6.3, Android detects *Screen Off* and *Screen On* events. On the other hand, iOS mobile software detects *Screen Locked* and *Screen Unlocked* events. Moreover, both platforms detect more events such as charging and discharging a battery. Nevertheless, I did not use these events.

Timestamp	UTC Time	Event	Battery level
Unix	Y-m-d H:M:OS	"Locked" "Unlocked"	$[0, 1]$

Timestamp	UTC Time	Event
Unix	Y-m-d H:M:OS	"Screen On" "Screen Off"

Table 6.3: Power state data format for iOS (1) and Android (2)

6.3.2 Data Imputation and Filtration

Data imputation or filtration was unnecessary as the Beiwe platform produced a reliable data set for each patient.

6.3.3 Features Extraction

1. Screen time

   The first feature that was calculated was screen time per day. This feature sums the duration between *Screen Unlocked* and *Screen Locked* event on iOS. On Android, a sum of durations between *Screen On* and *Screen Off* events is used. However, there are erased durations that take no more than 30 seconds. This filtration happens to erase the screen wake up, for example, due to notifications.

   $$\sum_{i}^{n} \Delta d_i \tag{6.9}$$

   where:

   Δd_i = duration between on and off events or unlocked locked events
   n = Total number of durations

2. Screen frequency

   The second feature that was calculated was screen frequency per day. This variable essentially can be described as the number of unlocks of a phone in a day.

6.4 Mood

6.4.1 Characteristics

As shown in Table 6.4, participants of the study were asked to report their mood daily in the Mindpax.me mobile application using energy, inner feeling and mood variable. Each variable had a continuous range from -1 to 1. The mood came in the form of CSV files with nested JSON files. The duration of the survey filling out was also recorded.

6.4.2 Data Imputation and Filtration

Firstly, a thorough analysis of data had to be done. There was a need to exclude patient *BA009* because mood values were not filled out correctly nearly every day. Each value was zero every day, which indicated a lack of discipline. Consequently, patients *BA008* and *BA005* were excluded because there was an excessively high percentage of missing data.

The percentage of missing values of the others participants was not high. 10 participants had complete data without a day missing. Others had a few days without data but never more than 20% per patient. The days with missing mood values were excluded.

6.4.3 Features Extracted

There was no need to extract or modify the majority of features.

1. Mood
2. Energy
3. Inner feeling
4. Duration of survey

Feature	**Value**
Mood	Scale from $[-1, 1]$ where -1 is the saddest and 1 is the happiest
Inner feeling	Scale $[-1, 1]$ where -1 is the most anxious and 1 is the most relaxed
Energy	Scale from $[-1, 1]$ where -1 is the least energetic and 1 is the most energetic
Duration of survey	Time in seconds

Table 6.4: Everyday survey

5. Emotional state

I would like to classify mood, energy and inner feeling altogether as these features are highly subjective variables. Two approaches are commonly used in terms of studies that focus on mood prediction using passive data.

Firstly, in some studies *PHQ-2* questionnaires are used for mood evaluation [PAR+18]. These questionnaires consist of two questions and should be able to identify depression. However, this is not the task that I need to fulfil, and variables mood, energy and inner feeling are not suitable for this issue.

Secondly, in several studies, mood assessment is based on the circumplex model of affect [OT16, LLLZ13]. This model suggests that emotions are distributed within two-dimensional circular space [Rus80]. On the x-axis in the original model, there is usually valence. However, this could be substituted by our mood variable. On the other hand, there is usually arousal on the y-axis. Our energy variable could easily substitute arousal. Consequently, four states (+ 1 neutral) could be identified as it is done in a study focusing on characterising emotional state using heartbeat [VCl+14]. The whole model can be seen in Figure 6.2. Moreover, similar approach was used in already mentioned study [LLLZ13].

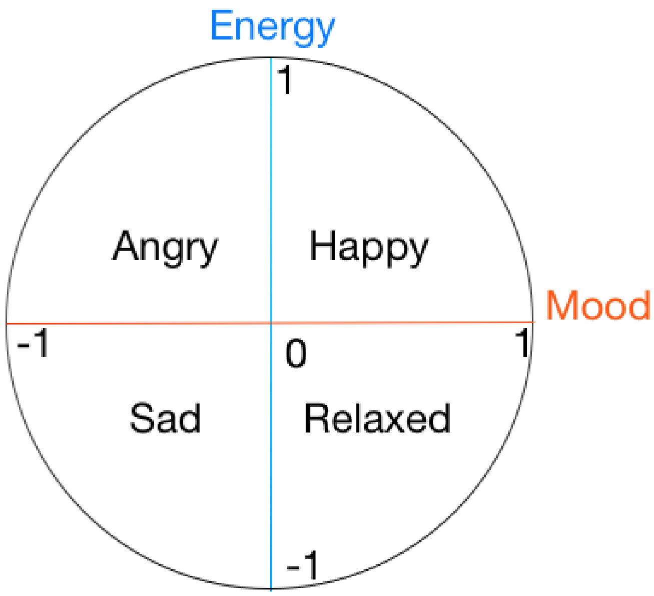


Figure 6.2: Edited circumplex model of affect

As shown in Table 6.5, the participant can be classified as angry, happy, sad, relaxed or neutral depending on the energy and mood values.

Emotion	**Condition**
Happy	Energy >0 AND Mood >0
Angry	Energy >0 AND Mood <0
Sad	Energy <0 AND Mood <0
Relaxed	Energy <0 AND Mood >0
Neutral	Energy $=0$ OR Mood $=0$

Table 6.5: Mood conditions

6.5 Actigraphy Data

6.5.1 Characteristics

The nature of actigraphy data were already described previously in Section 5.2. To sum up, the data came in the form of CSV files where each data row corresponds to the most significant magnitude vector of acceleration per 30 seconds.

6.5.2 Data Imputation and Filtration

The days with more than 10% of missing data were excluded. The data were quite robust. Thus, such a significant threshold was set. Moreover, the data were quite reliable, and the portion of missing data was usually due to the malfunctioned wristband. As shown in Table 6.6, this method filtered out just a few days from data sets of participants.

6.5.3 Features Extracted

1. Total sum

   This variable can be considered as the sum of total activity per day.

$$\sum_{i}^{n} s_i \tag{6.10}$$

   Where:

   s_i = value of magnitude of maximum vector per 30 seconds
   n = total number of observations in a day

Patient	Number of days before filtration	Number of days after filtration
BA001	28	28
BA002	28	27
BA003	28	28
BA004	22	21
BA005	25	21
BA006	28	28
BA007	28	28
BA008	25	22
BA009	28	27
BA0010	28	28
BA0011	27	19
BA0012	28	28
BA0013	28	27
BA0014	17	10
BA0015	28	27
BA0016	28	27
BA0017	28	28
BA0018	27	25
BA0019	28	27
BA0020	28	27

Table 6.6: Actigraphy: number of days before and after filtration

2. Peak hour sum

   This feature describes the biggest sum of the activity per hour in a day.

$$\max_{i}^{n} sum_i \tag{6.11}$$

   Where:

   sum_i = sum of actigraphy data in hour i
   n = 24 (hours in a day)

3. Peak hour

   This feature describes the exact hour with the biggest sum of the activity per hour in a day. In other words, it is the i from Equation 6.11.

4. Maximum median activity per hour

   This feature describes the highest median of medians of activity per hour.

$$\max_{i}^{n} med_i \tag{6.12}$$

   Where:

   med_i = median of actigraphy data in hour i in n (24) hours

6.6 Features Summary

All the calculated features for each patient can be seen in Table 6.7. These variables correspond to one day. The complete data set after the intersection of data from power state, mood, actigraphy and GPS data contains 15 patients with total 251 days of data with no missing values.

Feature	Origin
Total distance	GPS data
Location variance	GPS data
Number of clusters	GPS data
Time in each cluster	GPS data
Total time in clusters	GPS data
Entropy	GPS data
Normalised entropy	GPS data
Screen time	Power state data
Screen frequency	Power state data
Mood	Mood data
Inner feeling	Mood data
Energy	Mood data
Duration of survey	Mood data
Emotional state	Mood data
Total sum	Actigraphy data
Peak hour	Actigraphy data
Peak hour sum	Actigraphy data
Max median	Actigraphy data

Table 6.7: Summary of features calculated

Chapter 7

Exploratory Data Analysis

7.1 Introduction

In this chapter, I will focus on exploratory data analysis. I will look at the processed data set from the previous chapter. Moreover, I will take two points of view. Firstly, I will look at the data set of all patients as a whole unit, and secondly I will look at the patients individually. I will try to find meaningful correlations and relationships between variables in both cases. Moreover, studies do not often mention data and relationships which are not significant. Thus, I will also contemplate data that do not show notable values.

Nevertheless, I will have to consider the nature of this time as the whole world suffers from the COVID-19 pandemic. People's movement and mobility are lower than ever before in nearly every country [NBC^{+}21]. Moreover, some studies that focus on people's mental health during these times found simultaneously altered activity and worsened well-being [LN20]. These facts I have to take into account and validate the results accordingly.

7.2 General Based Approach

Looking at the data set as a whole, we can see in Figure 7.1 that mood values are generally higher. Thus, this fact confirms that we had mentally healthy patients in our data sample. On the other hand, as shown in Figure 7.2, *energy* values range on the whole interval $[-1, 1]$. The *inner feeling* can be considered, as shown in Figure 7.3, somewhere between the *mood* variable and *energy* feature. Considering our *emotion* variable described in Chapter 6 we can see that people were generally *happy* in our sample of people, as can be seen in Figure 7.4. Moreover, our reasoning can be confirmed by Figure 7.5 where *mood* reaches higher values and contains more outliers.

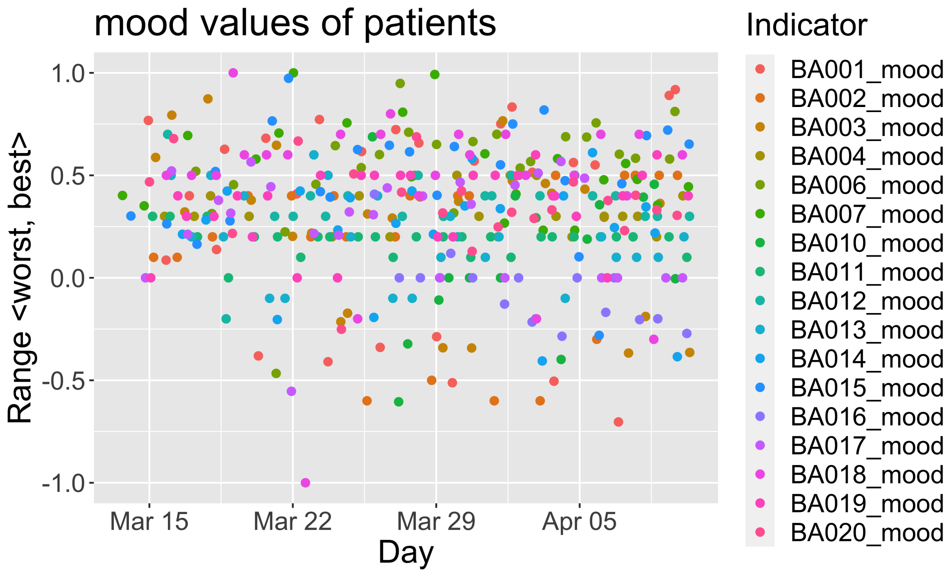


Figure 7.1: Mood: whole sample

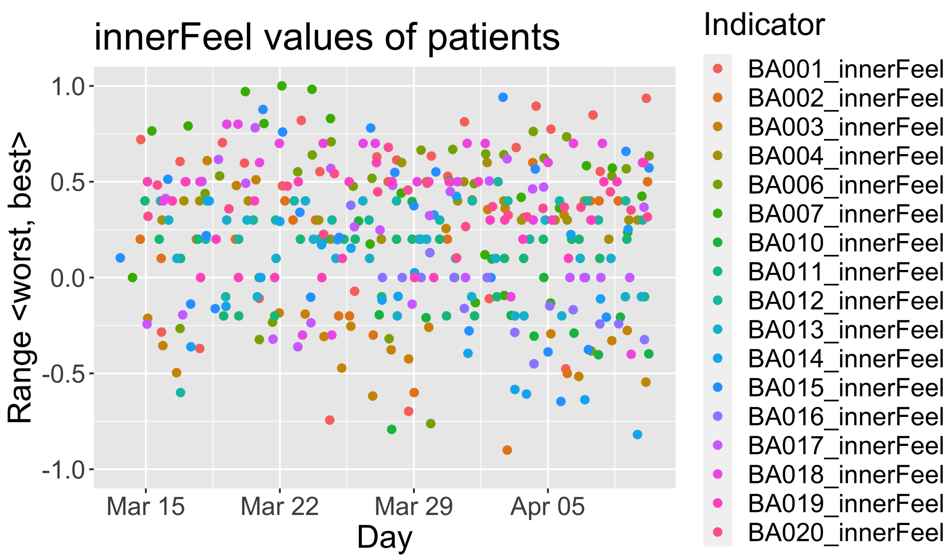


Figure 7.2: Inner feeling: whole sample

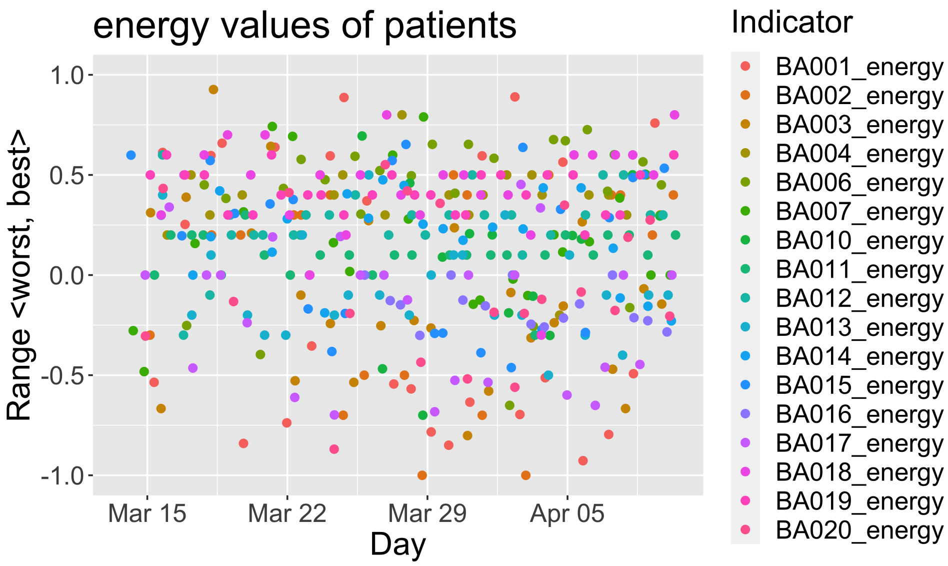


Figure 7.3: Energy: whole sample

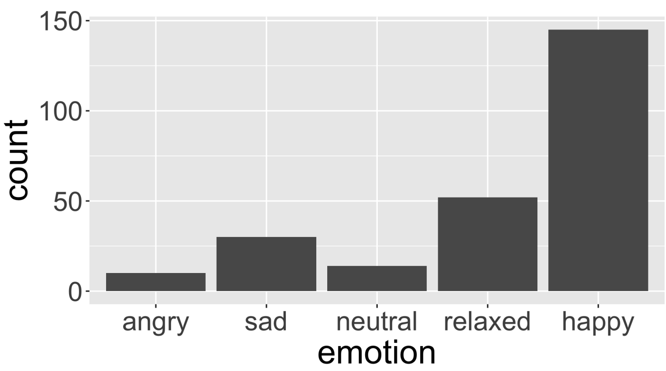
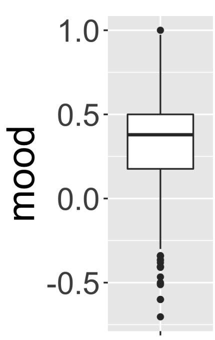


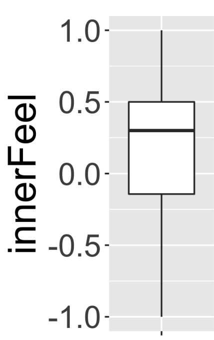
Figure 7.4: Emotions: whole sample

To contemplate features, I will use box plots. This technique is excellent for summarising data. The box plot uses median, approximate quartiles *(25th and 75th percentile)* and the lowest and highest data points to identify the level, spread and symmetry of data distribution. The interquartile range is the central 50% of data points or the area between the $75th$ and the $25th$ percentile of a distribution. Moreover, I will be using the term outlier. A point is an outlier if it is above the $75th$ or below the $25th$ percentile by a factor of 1.5 times the interquartile range. [WPK89]

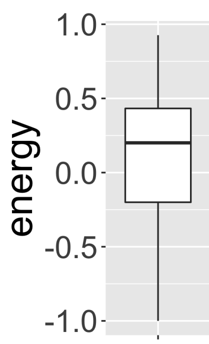
As shown in Figure 7.6, the *number of clusters* visited per day is relatively consistent across the sample with a few outlier data points. This finding might be due to the nature of this time as people are generally less dynamic and relatively stationary. This fact can also be validated in the box plot of the *total distance* travelled per day. Moreover, *time in clusters* or someone might say time spent not moving from related box plot, is relatively high too. A box plot of the *number of unlocks* of a phone can be seen in the same Figure 7.6. People are consistent across the sample with a few outliers that use phone excessively. This fact can be approved in the box plot of *screen time*. There consistent *screen time* is indicated with a few outliers. Interestingly, as shown in the exact figure in box plots of *peak hour*, it is concentrated more in the evenings.



(a) : Mood

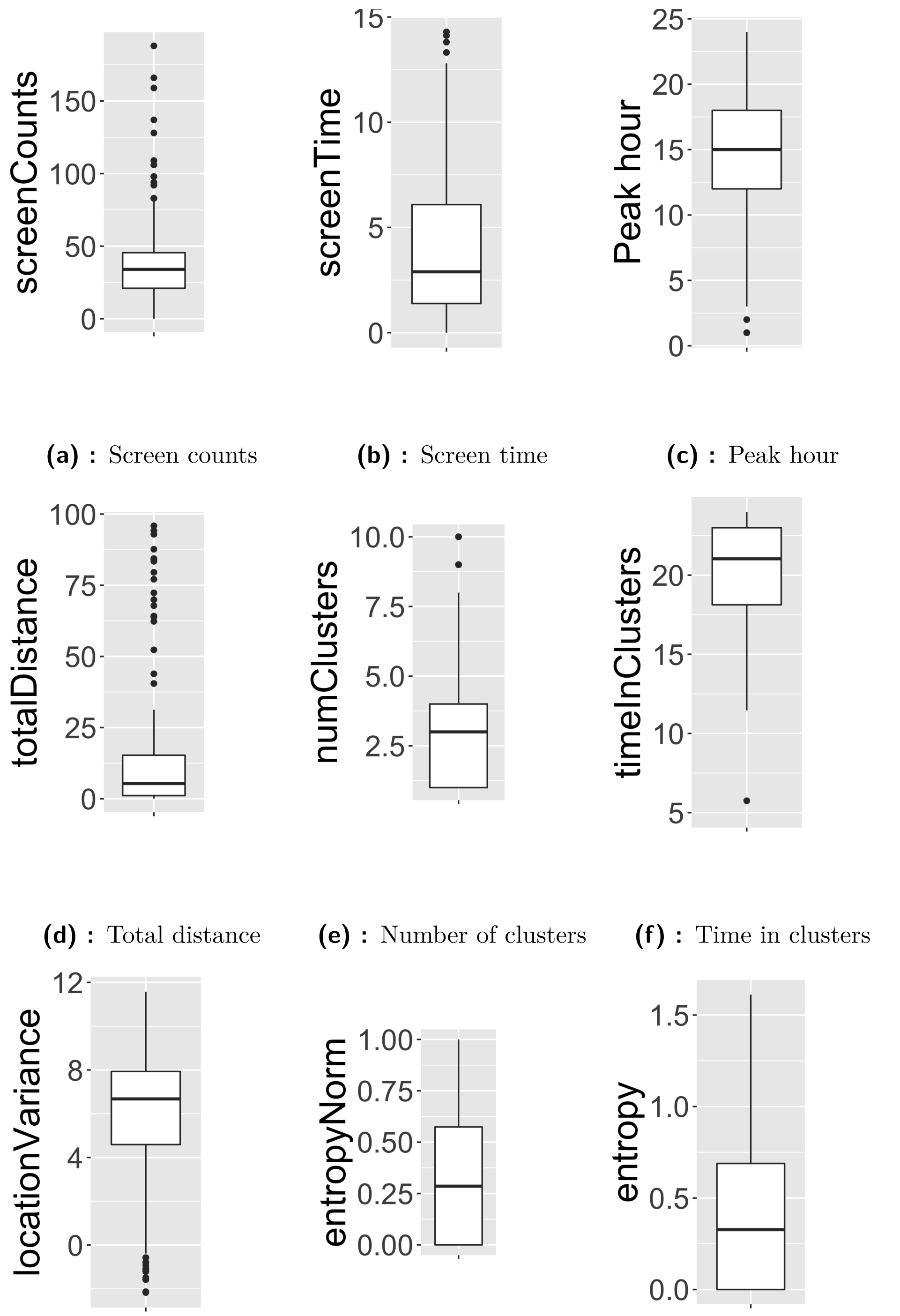


(b) : Inner feeling



(c) : Energy

Figure 7.5: Box plots whole sample part a



(a) : Screen counts **(b) :** Screen time **(c) :** Peak hour

(d) : Total distance **(e) :** Number of clusters **(f) :** Time in clusters

(h) : Entropy

(g) : Location variance **(i) :** Normalised entropy

Figure 7.6: Box plots whole sample part b

As a next step, correlation analysis was done. For this analysis, I used the Pearson correlation coefficient. This method measures a linear dependence between two variables. Also, the distribution of variables has to be normal. In our case, this is a valid prerequisite in most of our features. However, we have to consider that sometimes our data have a few outliers that could corrupt the correlation coefficient. The equation for computing correlation coefficient r is described below.

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(y - m_y)^2 \sum(x - m_x)^2}} \tag{7.1}$$

where:

m_x = mean of variable x
m_y = mean of variable y

Sometimes a strong correlation can be found, but it might happen due to an error. Thus, it is essential to conduct a statistical significance test. This test provides information on whether our correlation from the sample could also work in population. The significance test can be done with a hypothesis test.

The hypothesis test tells us if the population correlation coefficient is close to 0 or not. Also, we have two hypotheses. The *null hypothesis* says that the result happened due to a chance, and the *alternative hypothesis* says that the result did not happen due to a chance. Then *t-test* was done using the formula below.

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2} \tag{7.2}$$

where:

n = number of observations in x and y variable

Every *t-value* corresponds to a *p-value* representing the probability that the result happened due to a chance. We can compute degrees of freedom and then look up the *p-value* using the number of degrees of freedom and *t-value*. Moreover, the confidence interval is set to 95% in our case. The confidence interval implies a significance level which is 0.05. If the *p-value* is small than our significance level, we reject our null hypothesis. This rejection essentially tells us that the correlation is statistically significant. On the other hand, if the *p-value* is bigger than our significance level, we cannot reject the null hypothesis. Thus, we conclude that the correlation is not statistically significant.

Correlations computed for our data set can be seen in Figure 7.7. All the statistically insignificant correlations are blank. To start with the strongest correlation, we can see that actigraphy variables such as *sumActivity*, *maxSumActivityPerHour*, *maxMedianActivityPerHour* are all strongly correlated. This finding is interesting but not surprising. All the variables come from the same actigraphy source but are computed differently. Thus, this suggests that people with more activity have more activity in *peak hour* and have a high median probably from *peak hour*.

Looking at the *entropy*, one can see a strong correlation between *normalised entropy* and *entropy*. This conclusion is not surprising as the *normalised entropy* is based on raw *entropy*.

Let me focus on interesting correlations now. Looking at the *mood*, *energy* and *inner feeling* variables, we can see several interesting phenomena. Firstly, all three variables are positively correlated altogether. Secondly, *energy* and *mood* seem to grow with more clusters visited per day. This fact suggests that the more clusters are visited in a day by a person, the more energy or happiness he or she has and vice versa. Thirdly, *inner feeling* shows an interesting positive correlation between actigraphy data, for example, the *sum of activity*. This finding might suggest that with a higher activity comes a more relaxed inner feeling. *Mood* also correlates with the variables from actigraphy data in a similar way.

Screen time and *screen count* correlate with each other as with more unlocks of a phone comes higher *screen time*. However, screen count also correlates with variables such as *location variance* and *entropies*. It seems that the bigger the *location variance* and more distributed time across the clusters, the bigger the *screen count*. Interestingly, *screen count* positively correlates with a *sum of activity* per day.

Significantly, *time in clusters* negatively correlates with *entropy*. This surprising finding suggests that the more uniform distribution across the clusters is, the less time is spent in a stationary state per day, and more time is spent on the road. *Time in clusters* also positively correlates with the sum of activity which is rather suspicious as I would suggest the other way around.

Lastly, *total distance* positively correlates with a *number of clusters* which is logical as the more distance travelled relates to more visited clusters. It also negatively correlates with *time in clusters*, which is not surprising as the more time an individual spends travelling, the less time is spent in stationary states. Significantly, *total distance* positively correlates with *location variance* because of the fact that with immense distance comes more significant variance in movement. Moreover, the *number of clusters* is negatively correlated with *time spent in clusters*. Further, the bigger the *number of clusters*, the more significant the *location variance*. However, the last statements are logical since the number of clusters implies more time spent in motion and more vital variance in movement.

To sum up, some general patterns in the whole sample can be seen, but many correlations are erased due to the hypothesis testing. Moreover, we have to take into account that correlation is not a causal relationship, and thus we cannot say if x influences y, y influences x or another variable z influences both x and y.

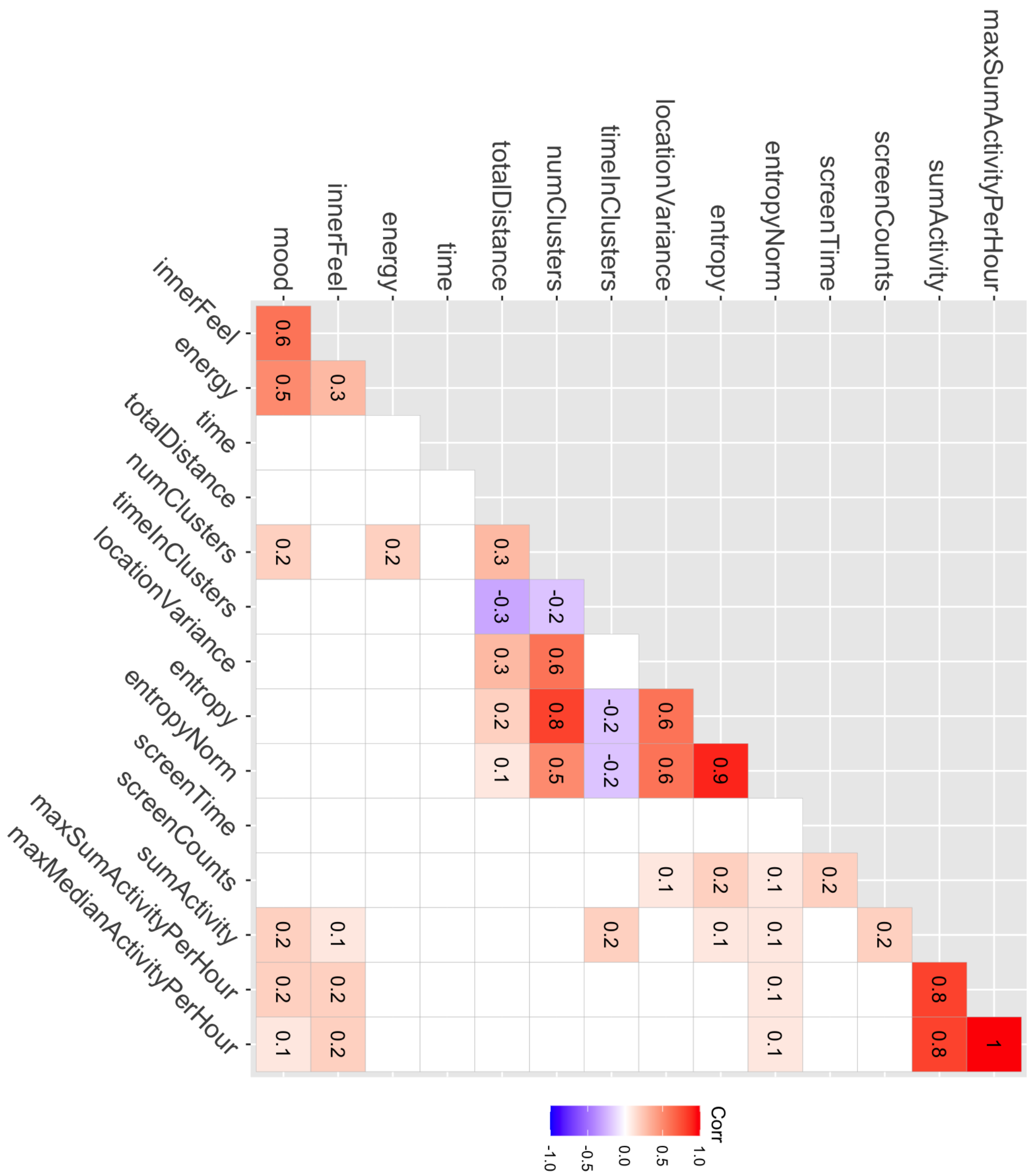


Figure 7.7: General approach: correlations

7.3 Individual Based Approach

Looking at each individual's data set, it can be easily said that each person behaves differently. In Figure 7.8 we can see the number of emotions felt by our patients. We can see that our sample generally feels *happy* with just a few points of *angry* emotion. Moreover, in Figure 7.9 we can conclude that the patients who use a more considerable portion of intervals in mood variables use the same approach in all three features *mood*, *inner feeling* and *energy*. In other words, the stability or instability is the same across the three variables.

As already described earlier, the majority of our sample does not travel that much and instead spends more time at home, as shown in Figure 7.10. However, I believe this is due to the restrictions applied because of the COVID-19 pandemic. Nevertheless, to be sure, it would be essential to have data from time without the COVID-19 pandemic.

As shown in figure 7.11, the activity is relatively consistent across the sample, and *peak hour* is oriented more to the evenings, which I find pretty interesting. Moreover, an interesting pattern can be seen in Figure 7.12, where one participant can open the mobile phone quite frequently but does not spend much time on it, and another patient can unlock the phone just a few times but spends an immense amount of time on it.

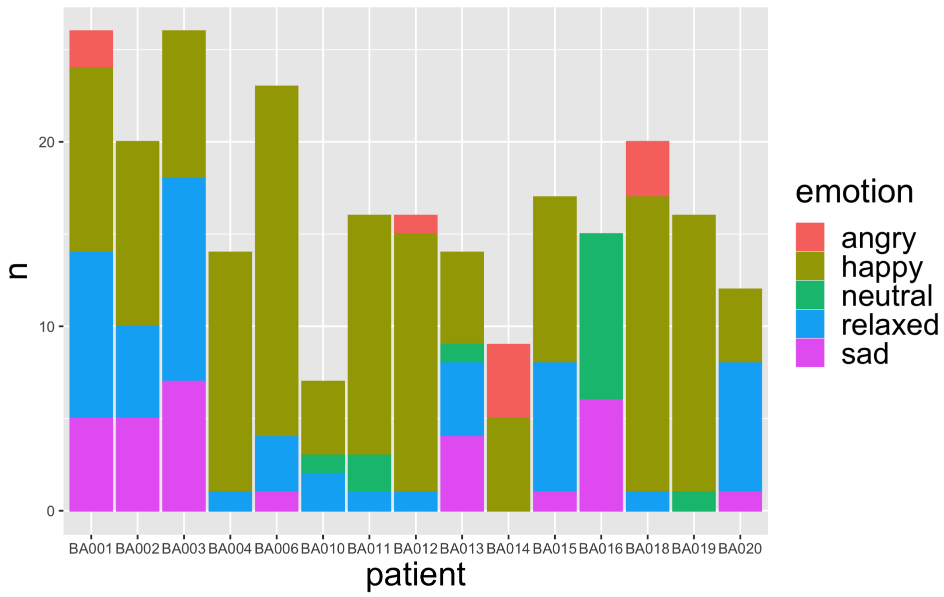


Figure 7.8: Emotion patients summary

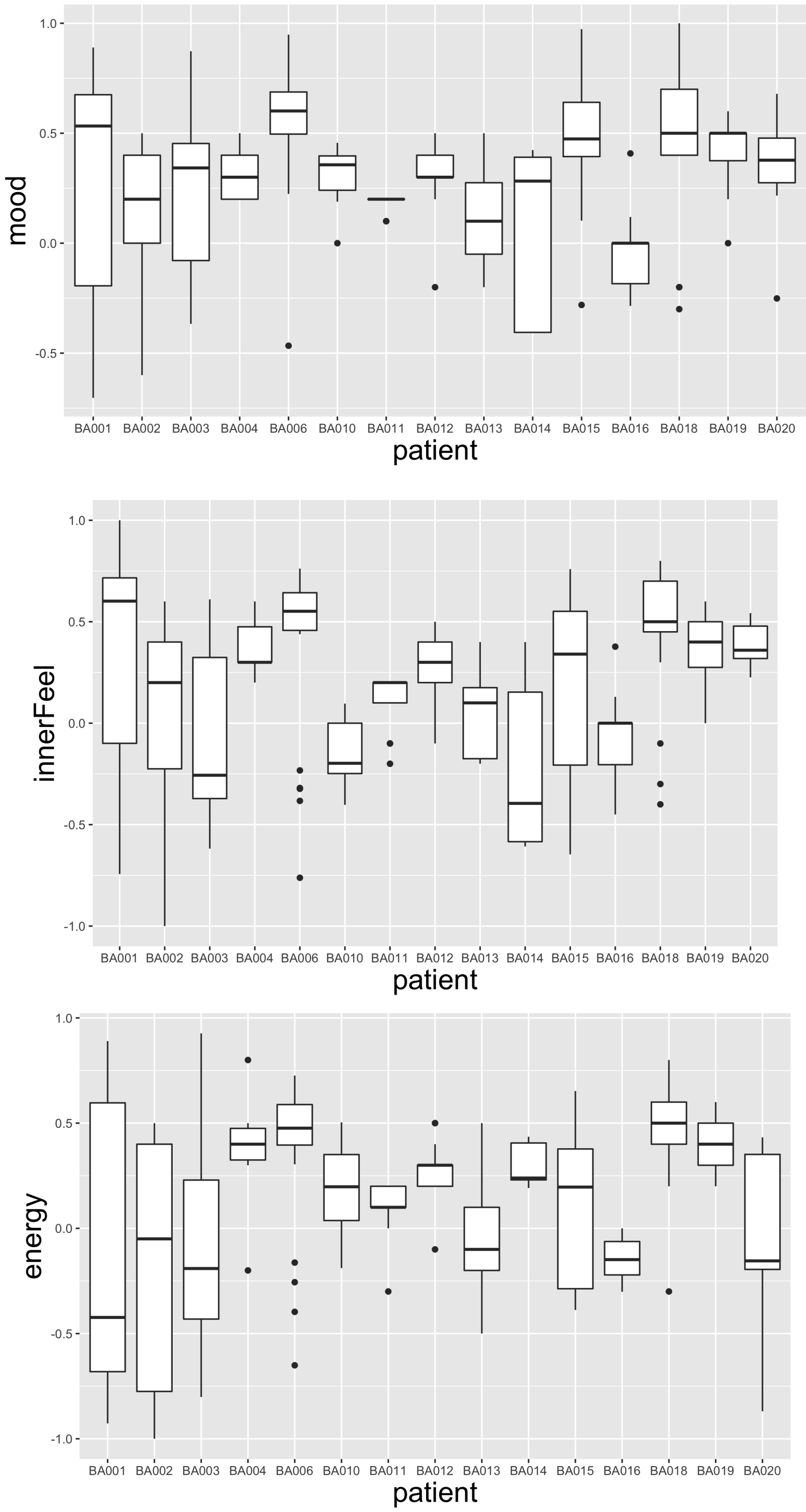


Figure 7.9: Mood, inner feeling and energy box plots

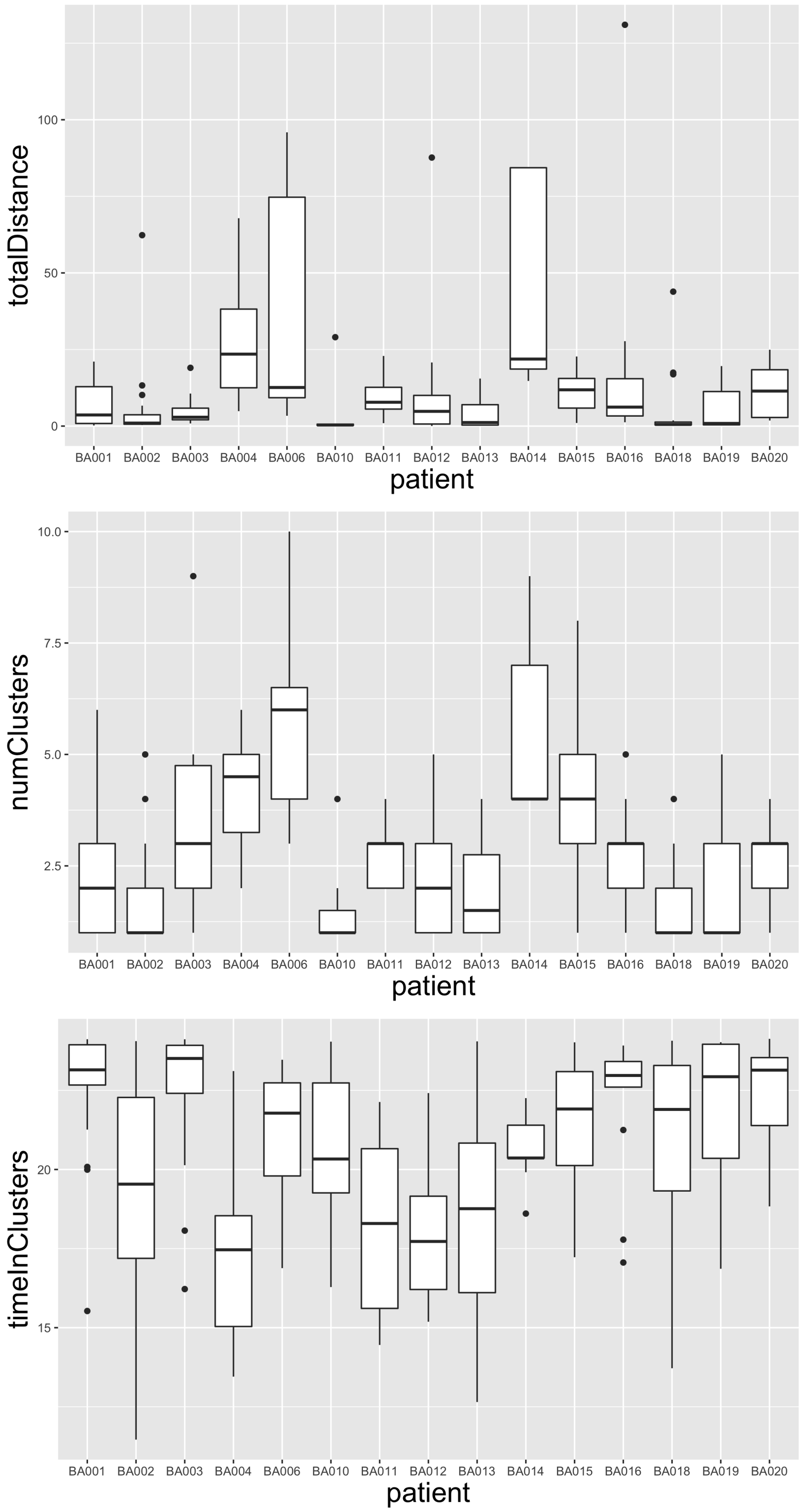


Figure 7.10: Total distance, number of clusters, time in clusters box plots

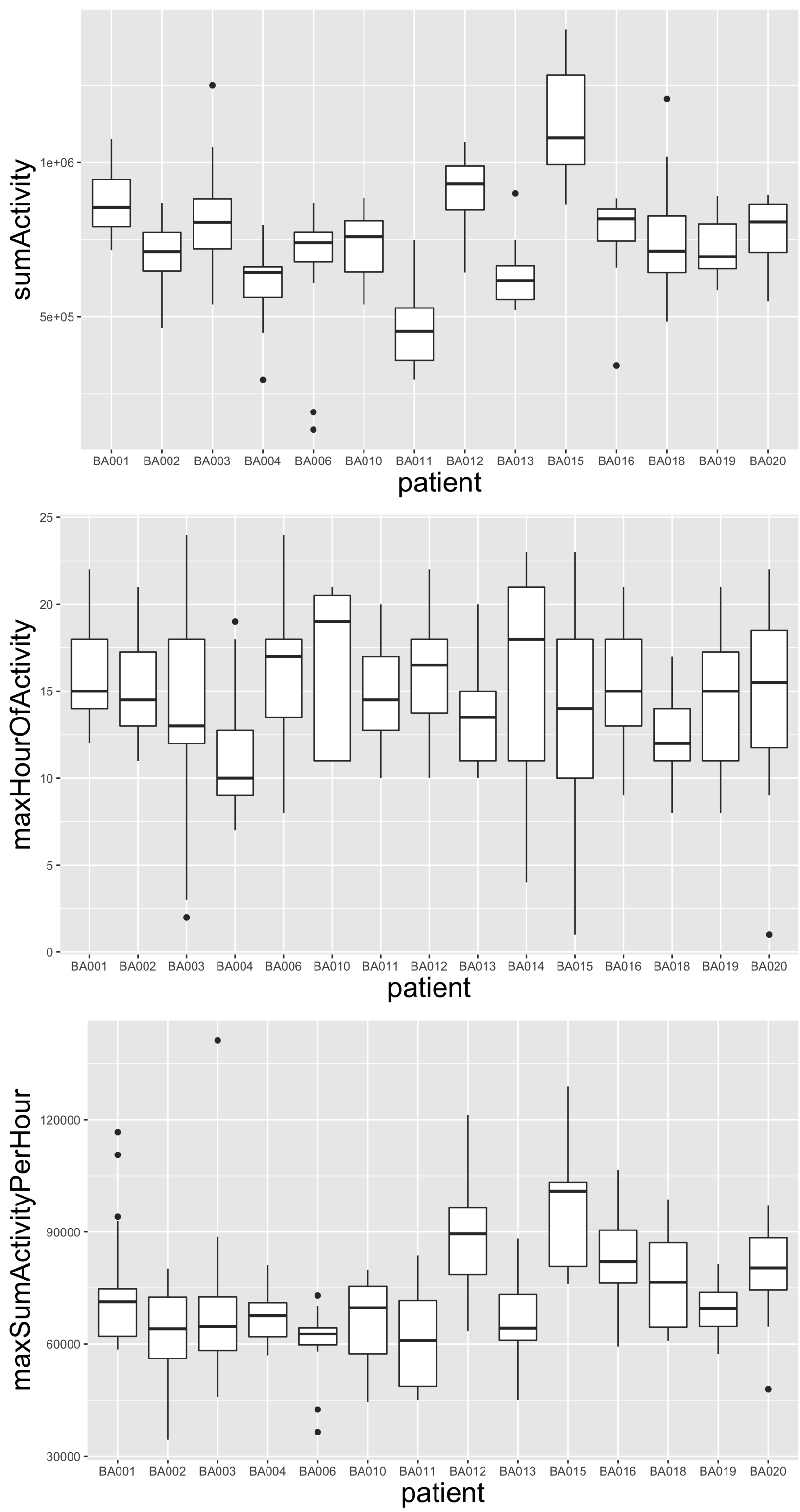


Figure 7.11: Total sum of activity, peak hour and peak hour sum box plots

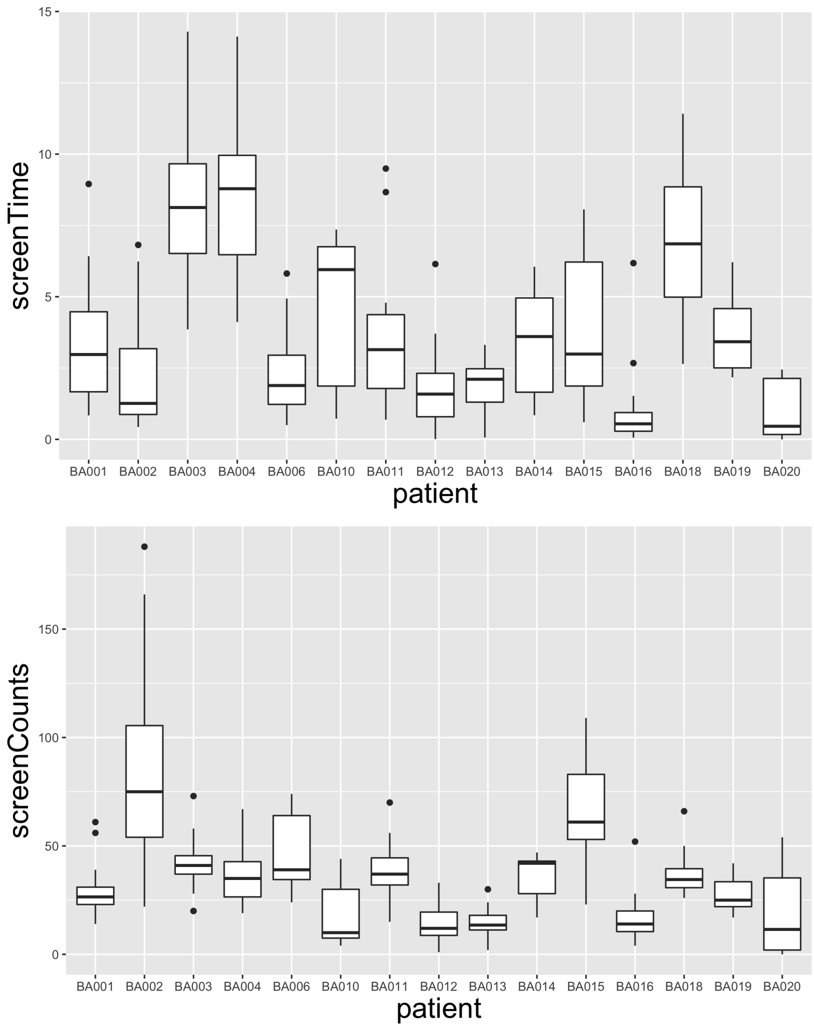


Figure 7.12: Screen time and screen count box plots

As shown in Figure 7.13, patient *BA002* has strong correlations across the *mood*, *energy* and *inner feeling* variable. Moreover, *screen time* and *screen count* is positively correlated with activity features such as the *sum of activity*. Interestingly, the *time* (duration of the survey) of filling out questionnaires is negatively correlated with activity features. On the other hand, in Figure 7.14, we can see a different set of correlations of patient *BA004* which are, again, quite strong. Interestingly, the *sum of activity* correlates with *inner feeling*. Further, time in a stationary state *(time in clusters)* correlates positively with *screen time* and negatively with the *sum of activity*. This finding suggests more time spent on a mobile phone when not moving and simultaneously a lower activity, which is the opposite from the first patient.

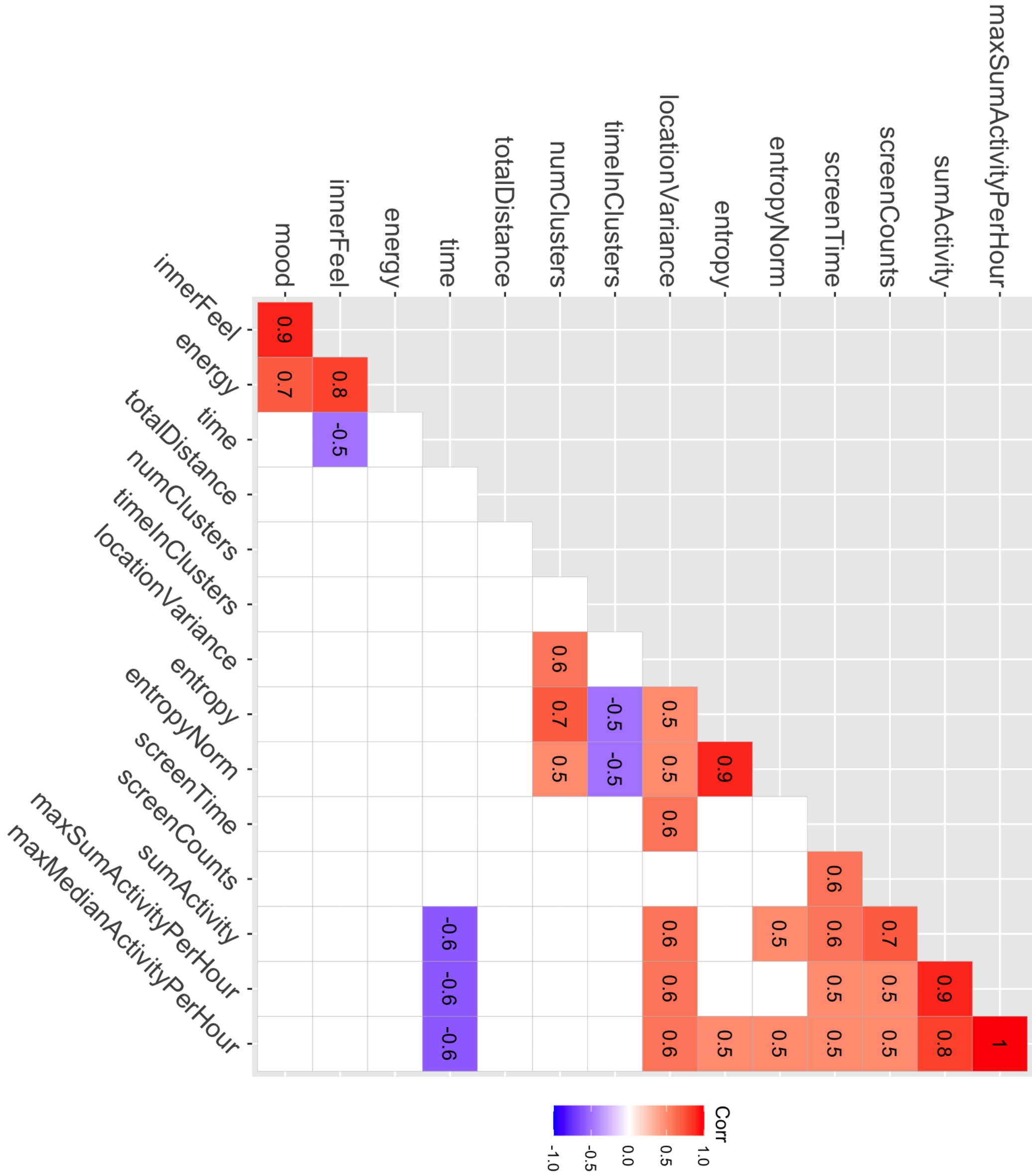


Figure 7.13: Individual approach: BA002

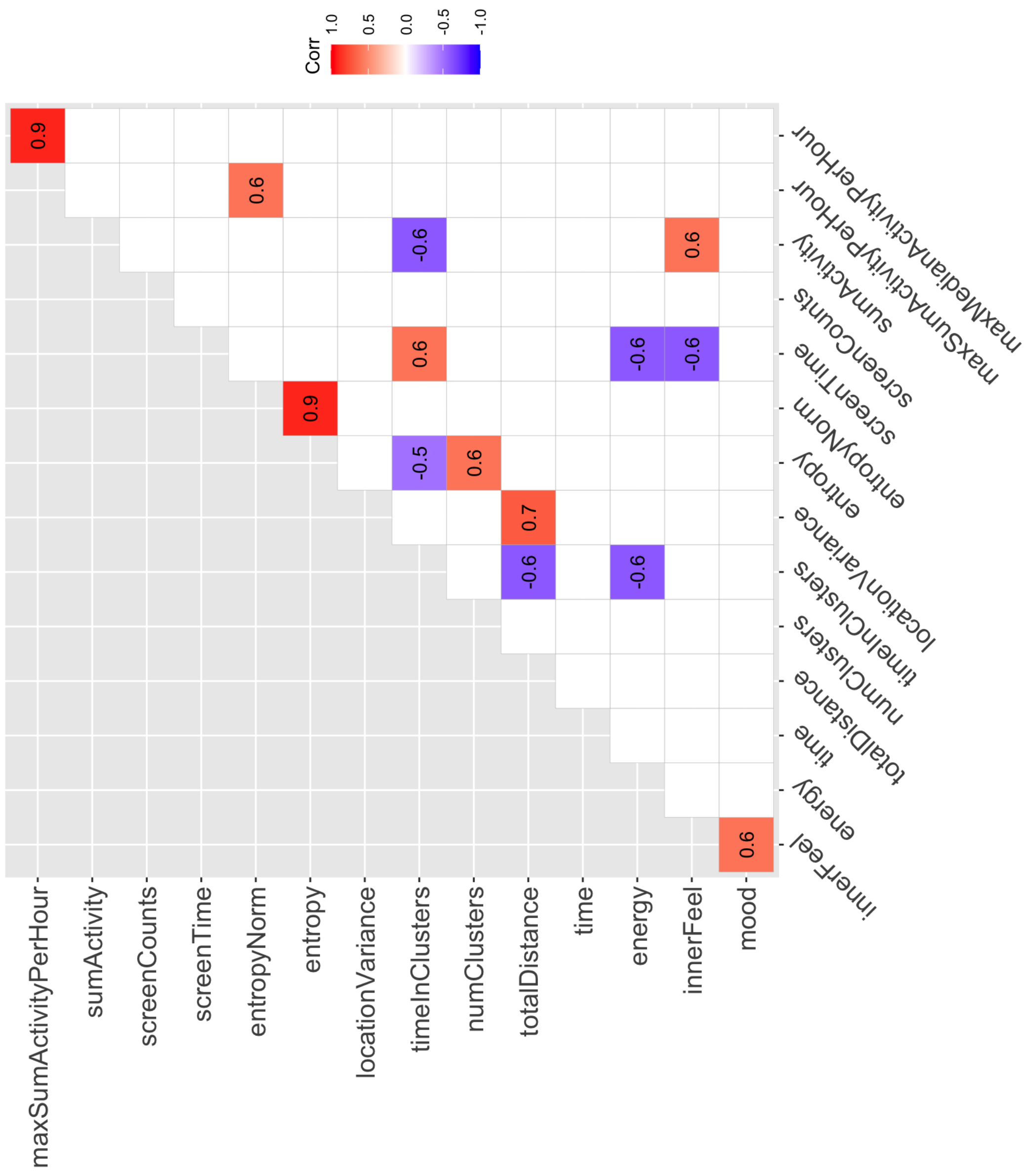


Figure 7.14: Individual approach: BA004

7.4 Conclusion

After analysing the whole sample and analysing participants individually, I can easily conclude that there is a high variability across the participants. This fact suggests that each person behaves differently and has different patterns and related data. A person is a complex entity with different habits, and this is what the data tells me. In the general approach, some weak patterns were found, but we have to consider the size of the sample and the fact that people spend more time at home nowadays. Moreover, our sample included healthy people. The variability could be even higher with people with mental health issues.

Chapter 8

Classification

8.1 Introduction

In this chapter, I will try to classify the emotional states described in Chapter 6. Moreover, I will present the state of the art of classification and prediction of mood. As described later, classifying or predicting mood is an arduous task with great potential but currently lacking accuracy and research [LLLZ13].

8.2 State of The Art

One might think that mood prediction, classification or detection using these data types is a profound and explored destination. Only a few studies focus on this topic and predominantly work with a small sample that makes the replication of the methods used in mood prediction or classification hard. Primarily these studies focus on classifying depression or depression severity. For example, it is believed that GPS and phone usage features might detect depressive symptom severity as depression is associated with behavioural movement components such as changes in sleep, reduction in activity or psychomotor retardation [SZK+15]. Another study supporting this statement focuses on GPS data for predicting depression among undergraduates [JC20].

Mood prediction or detection can be seen as a massive advantage in the business or advert industry. However, there is a risk that highly idiographic associations between passive features and daily mood might be learned when the ultimate goal is to find a generalised pattern over the whole population. [LLLZ13]

Moreover, some studies express concern over the connection of passive data streams of phone and the prediction of daily mood as variability in features between individuals in mood disorders overwhelms any predictive information in these features [PAR+18].

8.3 Random Forest

For the classification of the emotional states, I will use classification using the random forest. This supervised learning method is based on the *divide and conquer* approach. The main principle behind these ensemble methods is gathering weak learners and combining them into one strong learner [Zho09]. The random forest method itself is easily parallelisable, can deal with small data sets and is recognised for its accuracy [BS15].

A random forest classifier is a method that uses decision trees to make a classification. These trees are created by taking a subset of training data through replacement, known as the bagging approach. About two-thirds of the samples are utilised for training the trees, and the rest is used for internal cross-validation to identify the success of the random forest. The error used for this validation is called an *out-of-bag error*. Each tree is created independently, and the total number of trees depends on the chosen value. Moreover, each node is split by a user-defined number of features that are chosen randomly. This variable is usually called *mtry*. Consequently, when passing new data to the forest, each tree votes for exact classification and then the majority vote of classifications wins. [BD16]

I tried to classify the emotional state of a day with four categories *angry*, *happy*, *relaxed* and *sad*. Rows with the *neutral* emotional state were omitted, as their classification led to inconsistency in classification, and there were just a few occurrences. Moreover, as shown in Table 8.1, I used 11 features in my model from the data set described in Chapter 6 where each row corresponds to one day. To sum up, this data set contains 237 days - rows of data from 15 participants. However, I have to acknowledge that the sample was imbalanced as our study participants felt *happy* more than any other

Feature used
Total distance
Location variance
Number of clusters
Time in each cluster
Total time in clusters
Entropy
Normalised entropy
Screen time
Screen frequency
Peak hour
Peak hour sum
Total sum of activity

Table 8.1: Features used in random forest classification from Chapter 6

emotion. In detail, there were 145 rows with *happy* classification, 30 rows of *sad* classification, 52 rows of *relaxed* emotional state and just 10 data entries with *angry* classification.

Firstly, I divided this already small data set into *train* and *test set*, keeping in mind the data distribution and the data disjunction. The train set included 70% of data and the test set 30% of data. However, it was essential to balance the data set in order not to overfit the top category. In general, for tackling this issue, two main methods are used. *Undersampling* leads to ignoring a fraction of the data within the main category. The main disadvantage of undersampling is that it can discard meaningful data. On the other hand, *Upsampling* tries to create new or duplicates data samples to balance the data set and achieve the same number of rows within each category. The downside of this approach is that it makes overfitting more likely. [KKP05]

There are many modifications of undersampling and upsampling. However, I decided to apply upsampling due to the size of the train set. I did not want to make the data set even smaller. I applied the random upsampling approach, which tries to add data samples into the set by copying data of minority classes until it reaches balanced data set [KKP05]. Specifically, I applied the upsampling just on the train set. The test set was left without a change to represent the original distribution of data. Then, I ran a random forest algorithm with three times repeated 10-Fold cross-validation. This procedure was done because of the small data set and in order to tune parameters to generalise the model as much as possible. For tuning, I used grid search with a 5, 25, 50, 100, 250 and 500 number of trees *(ntree)* for the random forest and number of variables that the node can be split on at random *(mtry)*. These parameters can also be viewed in Figure 8.1.

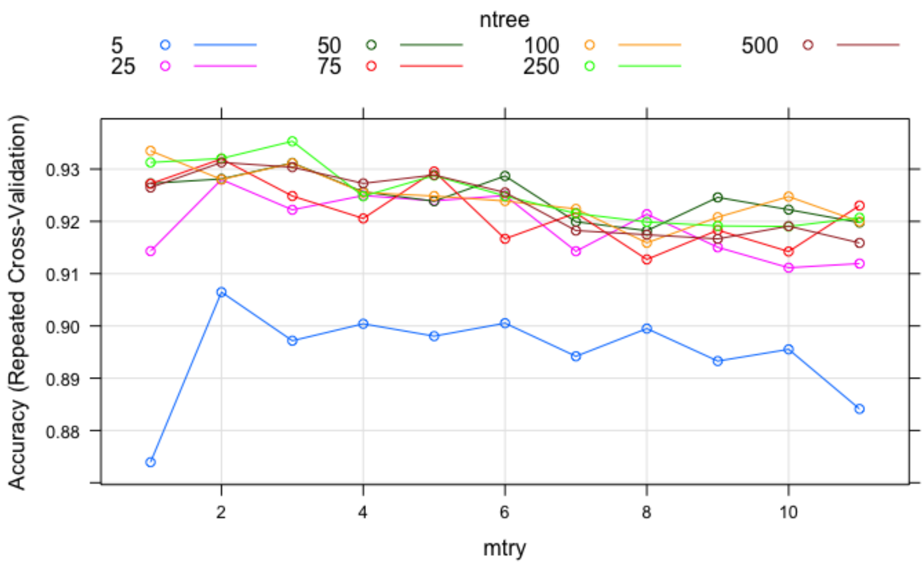


Figure 8.1: Random forest grid - search 3 times repeated 10-fold cross-validation on train set

Then, the final random forest model was chosen based on the best mean of accuracy after cross-validation. As seen in Figure 8.1, the greatest average of accuracy was with *mtry* set to 3 and the number of trees in the forest set to 250. Then the final model was used to classify the test set. The results, which I will elaborate on later, can be seen in Table 8.4. Moreover, a comparison of accuracy on training and test set is provided in Table 8.3.

8.4 Metrics and Results

For evaluating the performance of a model, I will use measures explained below. However, firstly, I would like to define a couple of terms. *True positives (TP)* are elements that are labelled as positive from the model and are actually positive. *False positives (FP)* are elements that are labelled as positive but are actually negative. On the other hand, *true negatives (TN)* are elements that were labelled as negative from the model and are genuinely negative. *False negatives (FN)* are elements that are labelled as negative but are positive. Moreover, all measurements are based on a paper describing metrics for multi-class classification [GBV20].

1. Precision

   Precision essentially describes how much we can trust the model when it predicts an individual as positive. In other words, it is the proportion of units that our model evaluates as positive and are actually positive.

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{8.1}$$

   where:

   k refers to exact class

2. Macro average precision

$$MacroAveragePrecision = \frac{\sum_{k=1}^{K} precision_k}{K} \tag{8.2}$$

   where:

   K = total number of classes

3. Weighted macro average precision

$$MacroAveragePrecisionW = \sum_{k=1}^{K} precision_k \cdot w_k \tag{8.3}$$

   where:

   w_k = (number of occurences of k in set)/(total number of points in set)

4. Recall

Recall measures the model's predictive accuracy for the positive class. In other words, it measures the ability to find all positive units in the data. Recall for each class is described below.

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{8.4}$$

where:

k	refers to exact class

5. Macro average recall

$$MacroAverageRecall = \frac{\sum_{k=1}^{K} recall_k}{K} \tag{8.5}$$

where:

K	= total number of classes

6. Weighted macro average recall

$$MacroAverageRecallW = \sum_{k=1}^{K} recall_k \cdot w_k \tag{8.6}$$

where:

K	= total number of classes
w_k	= weight of class k

7. F1-score

$$F1score_k = 2 \cdot \frac{precision_k \cdot recall_k}{precision_k + recall_k} \tag{8.7}$$

where:

k	refers to exact class

8. Macro f1-score

$$F1scoreM = \frac{1}{K} \cdot \sum_{k=1}^{K} F1score_k \tag{8.8}$$

where:

K	= total number of classes

9. Macro f1-score weighted

$$F1scoreMW = \sum_{k=1}^{K} F1score_k \cdot w_k \tag{8.9}$$

where:

K = total number of classes
w_k = weight of class k

10. Balanced accuracy

Balanced accuracy is essentially an average of recalls. As explained earlier, recall tells how likely will an individual be classified correctly. Thus, this measurement deals with an average of this concept across the different classes.

Considering the results shown in Table 8.3 it has to be said that my model tends to overfit. This issue happened due to the upsampling method used to balance the train set. Possibly, a different upsampling method or combination of downsampling and upsampling method with my data set could be used. Table 8.2 shows the confusion matrix of classifications where rows correspond to prediction and columns correspond to reference values. As shown in Table 8.4, both precision and recall are high in the *happy* class. This fact suggests that the classifier is quite sure about the *happy* emotion and can detect nearly all *happy* emotions in the test set. On the other hand, when it classifies emotion as *sad*, it is not that confident and detects just a third of *sad* occurrences in the test set. Interestingly, when it classifies emotion as *angry*, we can be pretty confident that the emotion is undoubtedly *angry*. However, the model detects just half of these emotions in the test set and we have to acknowledge the small amount of *angry* data. When looking at results of *relaxed* emotion, the precision is second best, but the classifier detects just two-thirds of *relaxed* emotions in the test set.

If all the classifications on test set were labelled as *happy*, the accuracy of this model would be 0.63. Thus, our accuracy 0.79 on the test set in Table 8.3 is slightly better than homogenous classification.

Reference **Predicted**	Angry	Happy	Relaxed	Sad
Angry	1	0	0	0
Happy	1	43	5	3
Relaxed	0	1	10	3
Sad	0	1	1	3

Table 8.2: Confusion matrix on test set: predicted - rows, reference - columns

	Train set	Test set
Accuracy	1	0.79
Balanced accuracy	1	0.61

Table 8.3: Train:test split result

	Precision	Recall	F1-score
Angry	1	0.5	0.67
Happy	0.82	0.96	0.88
Relaxed	0.71	0.63	0.67
Sad	0.6	0.33	0.4
macro average	0.78	0.61	0.66
macro average weighted	0.79	0.85	0.81

Table 8.4: Classification score random forest: test set

8.5 Discussion

One might suggest that the results of the random forest model mentioned earlier are a great success. However, I would be somewhat cautious. The main problem of the model set is the lack of data. I believe it is not general enough and tends to overfit. With more data, the model could be more general and could learn patterns across the emotions with fewer occurrences. Further, fifteen patients enrolled in the study were not a general enough population sample. The essential part might be to test the model on unseen participants with totally different behaviour patterns. However, this was not possible as our data sample was homogenous. Moreover, the data acquired are heavily based on the activity and dynamic of participants. To remind the data were gathered during the COVID-19 pandemic and thus, it might classify the same participants during "normal" regime inconsistently.

From a model perspective, there is a broad range of opportunities to make classification better. Firstly, different supervised learning models could be tried and used. An interesting approach might be to use deep learning models as it is a relatively unexplored destination of digital phenotyping [MSMM17]. Secondly, different features with the data I acquired during the study could be calculated. Thirdly, a better balancing method could be used that reduces overfitting.

Chapter 9

Conclusion

This thesis investigated a new phenomenon called digital phenotyping. The significance of this method, current state of the art and thorough explanation were provided in Chapter 2. Moreover, the challenges regarding data collection and data privacy were also discussed. Nevertheless, a vast potential of digital phenotyping for monitoring medical patients and aid in treating mental illnesses was described.

For thorough analysis, it was essential to collect data. Thus, in Chapter 3, current mobile data collection platforms and the data they can acquire were described. As it turned out, there were just two complete data collection platforms, working, available and used among researchers.

Further, a detailed comparison of these two platforms was given in Chapter 4. This comparison was made by deploying Mindlamp and Beiwe platforms and creating a pilot study. In this pilot study, 6 participants collected data simultaneously on both data collection platforms on both Android and iOS devices for two weeks. Consequently, a thorough comparison of data collection ability and of data types that these platforms can collect was created, and if necessary, the data were visualised. In this comparison, a missing ratio of data or similar problems were examined. However, it was essential to look at the comparison from different points of views:

1. The platform was examined from the researcher point of view. For example, price, deployment and eventual preprocessing were assessed.

2. User experience and user interface were compared, considering possible patients with medical issues.

3. Current and future development were analysed, and the final decision on which platform is currently better was made.

After the comparison and considering all points of view, it was decided that the Beiwe platform is currently a better option for digital phenotyping. Even though it is an older platform that lacks agile development, it is stable, does not tend to have data outages and can be used without obstacles or further maintenance. Still, it has its flaws, such as lack of data types collected, dependence on Amazon Web Services, or hosting price. However, the ability to edit the sampling frequency of sensors, small battery consumption or stability on the subjects' phone outperforms the Mindlamp platform immediately.

After finding out which platform was currently better, another pilot study was made. In Chapter 5, the pilot study was described considering all the ethical issues. The pilot study consisted of twenty people collecting data using the Beiwe platform for two months. Moreover, the participants filled out mood questionnaires each day and wore a band that collected actigraphy data. Thus, actigraphy as another phenomenon was also introduced and described in this chapter.

Consequently, a thorough analysis and preprocessing of a newly acquired data set was needed. These procedures were described in Chapter 6, where data specifically from the GPS sensor, the power state of the mobile phone, actigraphy band, and questionnaires were assessed in detail and preprocessed in order to have a stable data set. Furthermore, features from preprocessed data were extracted and combined into the final data set. An exciting part was to classify peoples' emotions. This classification was done by the circumplex model of affect, described in that chapter.

The final data set, where one row corresponds to one day, consisted at the end of 251 days. The data came from 15 participants, as some had to be excluded due to the incomplete or invalid data. The data set was examined in Chapter 7, together with visualisations providing a meaningful exploration of the data. Specifically, the data set was assessed from two points of view. Firstly, there was an examination of the data set as a whole sample. Secondly, the data set of each participant was examined, and a comparison to other participants was made using Pearson's correlation coefficient, box plots and other visualisations.

The conclusion presented that each participant had different behaviour patterns and different correlations between features. Moreover, the data were characterised by the time they were gathered. By this, I mean it was evident that people are less dynamic and are not that active as it was a lockdown in the Czech Republic because of the COVID-19 pandemic during data collection. Moreover, it was noted that participants felt happy in our data sample most of the days. This finding was expected as our data set contained mentally healthy people.

Consequently, in Chapter 8, a random forest classification was applied. The model was trained on 70% of data modified by upsampling using 3 times repeated *10-cross-validation* where grid search for hyperparameters tuning was used. After this procedure, a final model with 250 trees and 3 features considered at each node split was used. Then the test set was classified using this model. The accuracy on the test set was 0.79, which can be considered a mild success. Specifically, there is a massive success in classifying happy emotion, together with immense precision on angry emotion. Nevertheless, classifying sad emotion is not that excellent, and the classifier has problems with that emotion. A classification of relaxed emotion can be considered a mild success. However, new data of people with different behavioural patterns could give a new perspective to the model and make it more general.

Appendix A

Index

Appendix B

Bibliography

[AVN26] Jiri Anyz, Pavel Vostatek, and Daniel Novak, *Comparison of the mindg wristband with the motionwatch wristband for monitoring of physical activity and sleep*, https://www.mindpax.me/assets/docs/Validation_study_appendix.pdf, Accessed: 2021-04-26.

[BD16] Mariana Belgiu and Lucian Drguţ, *Random forest in remote sensing: A review of applications and future directions*, ISPRS Journal of Photogrammetry and Remote Sensing **114** (2016), 24–31.

[BHI+19] Elisa Brietzke, Emily Hawken, Maia Idzikowski, Janice Pong, Sidney Kennedy, and Claudio Soares, *Integrating digital phenotyping in clinical characterization of individuals with mood disorders*, Neuroscience Biobehavioral Reviews **104** (2019).

[BHO+10] Jan Berle, Erik Hauge, Ketil Oedegaard, Fred Holsten, and Ole Bernt Fasmer, *Actigraphic registration of motor activity reveals a more structured behavioural pattern in schizophrenia than in major depression*, BMC research notes **3** (2010), 149.

[BMS+13] Christopher Burton, Brian McKinstry, Aurora Szentagotai Tătar, Antoni Serrano-Blanco, Claudia Pagliari, and Maria Wolters, *Activity monitoring in patients with depression: A systematic review*, Journal of Affective Disorders **145** (2013), no. 1, 21–28.

[BPC+19] James Berry, Sabrina Paganoni, Kenzie Carlson, Katherine Burke, Harli Weber, Patrick Staples, Joel Salinas, James Chan, Jordan Green, Kathryn Connaghan, Josh Barback, and Jukka Onnela, *Design and results of a smartphonebased digital phenotyping study to quantify als progression*, Annals of Clinical and Translational Neurology **6** (2019).

[BS15] Gérard Biau and Erwan Scornet, *A random forest guided tour*, TEST **25** (2015).

[BTS+18] Ian Barnett, John Torous, Patrick Staples, Luis Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela, *Relapse prediction in schizophrenia through digital phenotyping: a pilot study*, Neuropsychopharmacology **43** (2018).

[BZSB+14] Dror Ben-Zeev, Stephen Schueller, Mark Begale, Jennifer Duffecy, John Kane, and David Mohr, *Strategies for mhealth research: Lessons from 3 mobile intervention studies*, Administration and policy in mental health **42** (2014).

[CBOS19] David Cote, Ian Barnett, Jukka-Pekka Onnela, and Timothy Smith, *Digital phenotyping in patients with spine disease: A novel approach to quantifying mobility and quality of life*, World Neurosurgery **126** (2019).

[Cen19] Ericsson Research Center, *Enabling internet for all*, `https://www.ericsson.com/en/mobility-report/articles/enabling-internet-for-all-wef-edition/`, 01 2019.

[CGP+19] Kathryn Connaghan, Jordan Green, Sabrina Paganoni, James Chan, Harli Weber, Ella Collins, Brian Richburg, Marziye Eshghi, J.P. Onnela, and James Berry, *Use of beiwe smartphone app to identify and track speech decline in amyotrophic lateral sclerosis (als)*, 09 2019, pp. 4504–4508.

[CSL+20] Alex Cohen, Elana Schwartz, Thanh Le, Tovah Cowan, Christopher Cox, Raymond Tucker, Peter Foltz, Terje Holmlund, and Brita Elvevåg, *Validating digital phenotyping technologies for clinical use: the critical importance of "resolution"*, World Psychiatry **19** (2020), 114–115.

[CZZ+17] Bokai Cao, Lei Zheng, Chenwei Zhang, Philip Yu, Andrea Piscitello, John Zulueta, Olusola Ajilore, Kelly Ryan, and Alex Leow, *Deepmood: Modeling mobile phone typing dynamics for mood detection*, 08 2017, pp. 747–755.

[DIJ13] Nikolaos Dimotakis, Remus Ilies, and Timothy Judge, *Experience sampling methodology*, 01 2013.

[ESM+18] Johannes Eichstaedt, Robert Smith, Raina Merchant, Lyle Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David Asch, and H. Schwartz, *Facebook language predicts depression in medical records*, Proceedings of the National Academy of Sciences **115** (2018), 201802331.

[FJBF+16] M Faurholt-Jepsen, J Busk, Mads Frost, Maj Vinberg, Ellen Christensen, Ole Winther, Jakob Bardram, and Lars Kessing, *Voice analysis as an objective state marker in bipolar disorder*, Translational Psychiatry **6** (2016), e856.

[FMH+11] Mads Frost, Gabriela Marcu, Rene Hansen, Karoly Szaántó, and Jakob Bardram, *The monarca self-assessment system: Persuasive personal monitoring for bipolar patients*, 01 2011, pp. 204–205.

[GBV20] Margherita Grandini, Enrico Bagli, and Giorgio Visani, *Metrics for multi-class classification: an overview*, 2020.

[GOB+14] Agnes Grünerbl, Venet Osmani, Gernot Bahle, J.C. Carrasco, S. Oehler, Oscar Mayora, C. Haring, and Paul Lukowicz, *Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients*, Proceedings of the 5th Augmented Human International Conference (2014).

[GPT+13] Andrea Gaggioli, Giovanni Pioggia, Gennaro Tartarisco, Giovanni Baldus, Daniele Corda, Pietro Cipresso, and Giuseppe Riva, *A mobile data collection platform for mental health research*, Personal and Ubiquitous Computing **17** (2013), 241–251.

[HTL19] Kit Huckvale, John Torous, and Mark Larsen, *Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation*, JAMA Network Open **2** (2019), e192542.

[Ins18] Thomas Insel, *Digital phenotyping: a global tool for psychiatry*, World Psychiatry **17** (2018), 276–277.

[Jar20] Angur Jarman, *Hierarchical cluster analysis: Comparison of single linkage,complete linkage, average linkage and centroid linkage method.*

[JC20] Nicholas Jacobson and Yeon Chung, *Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones*, Sensors **20** (2020).

[JHE05] Steven Jones, Dougal Hare, and Kate Evershed, *Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder*, Bipolar disorders **7** (2005), 176–86.

[JSW20] Nicholas Jacobson, Berta Summers, and Sabine Wilhelm, *Digital biomarkers of social anxiety symptom severity: Digital phenotyping using passive smartphone sensors*, Journal of Medical Internet Research **22** (2020).

[KCK+21] Mathew Kiang, Jarvis Chen, Nancy Krieger, Caroline Buckee, Monica Alexander, Justin Baker, Randy Buckner, Garth Coombs, Janet Rich-Edwards, and Jukka-Pekka Onnela, *Sociodemographic characteristics of missing data in digital phenotyping.*

[KKP05] Sotiris Kotsiantis, D. Kanellopoulos, and P. Pintelas, *Handling imbalanced datasets: A review*, GESTS International Transactions on Computer Science and Engineering **30** (2005), 25–36.

[LC14] Reed Larson and Mihaly Csikszentmihalyi, *The experience sampling method*, pp. 21–34, Springer Netherlands, April 2014 (English (US)).

[LLLZ13] Robert Likamwa, Yunxin Liu, Nicholas Lane, and Lin Zhong, *Moodscope: Building a mood sensor from smartphone usage patterns*, 06 2013.

[LN20] Iris Lesser and Carl Nienhuis, *The impact of covid-19 on physical activity behavior and well-being of canadians*, International Journal of Environmental Research and Public Health **17** (2020), 3899.

[LSW+97] Y Lecrubier, DV Sheehan, E Weiller, P Amorim, I Bonora, K Harnett Sheehan, J Janavs, and GC Dunbar, *The mini international neuropsychiatric interview (mini). a short diagnostic structured interview: reliability and validity according to the cidi*, European Psychiatry **12** (1997), no. 5, 224–231.

[LXZ19] Yunji Liang, Zheng Xiaolong, and Daniel Dajun Zeng, *A survey on big data-driven digital phenotyping of mental health*, Information Fusion **52** (2019).

[MBM+14] Alexander Markowetz, Konrad Błaszkiewicz, Christian Montag, Christina Switala, and Thomas Schlaepfer, *Psycho-informatics: Big data shaping modern psychometrics*, Medical hypotheses **82** (2014).

[MGMP17] Melvin Mcinnis, John Gideon, and Emily Mower Provost, *Digital phenotyping in bipolar disorder*, European Neuropsychopharmacology **27** (2017), S440.

[MHVB13] Yves-Alexandre Montjoye, Cesar Hidalgo, Michel Verleysen, and Vincent Blondel, *Unique in the crowd: The privacy bounds of human mobility*, Scientific reports **3** (2013), 1376.

[MQRP13] Yves-Alexandre Montjoye, Jordi Quoidbach, Florent Robic, and Alex Pentland, *Predicting personality using novel mobile phone-based metrics*, 04 2013, pp. 48–55.

[MSH20] David Mohr, Katie Shilton, and Matthew Hotopf, *Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age*, npj Digital Medicine **3** (2020).

[MSMM17] Gatis Mikelsons, Matthew Smith, Abhinav Mehrotra, and Mirco Musolesi, *Towards deep learning models for psychological state* *prediction using smartphone data: Challenges and opportunities*, 11 2017.

[MW18] Lisa Marsch and Andrew Wallace, *Opportunities and needs in digital phenotyping*, Neuropsychopharmacology **43** (2018), 1.

[NBC+21] Pierre Nouvellet, Sangeeta Bhatia, Anne Cori, Kylie Ainslie, Marc Baguelin, Samir Bhatt, Adhiratha Boonyasiri, Nicholas Brazeau, Lorenzo Cattarino, Laura Cooper, Helen Coupland, Zulma M. Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Bimandra Djaafara, Ilaria Dorigatti, Oliver Eales, Sabine van Elsland, Fabrícia Nascimento, and Christl Donnelly, *Reduction in mobility and covid-19 transmission*, Nature Communications **12** (2021).

[New18] Newswire, *iphone users spend* 101$ *every month on tech purchases, nearly double of android users, according to a survey conducted by slickdeals*, `https://prn.to/3u8aoYZ`, 10 2018.

[Nis20] Melanie Nisenson, *A smartphone application to investigate the relationship between digital media use and mental health using ecological momentary assessment in a clinical sample of youth: a feasibility study*, 2020.

[O'D21] Simon O'Dea, *Mobile os market share 2021*, `https://bit.ly/2PILTCE/`, 02 2021.

[Onn20] Jukka-Pekka Onnela, *Opportunities and challenges in the collection and analysis of digital phenotyping data*, Neuropsychopharmacology **46** (2020), 1–11.

[OR16] Jukka-Pekka Onnela and Scott Rauch, *Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health*, Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology **41** (2016).

[OT16] Andreas Olsen and Jim Torresen, *Smartphone accelerometer data used for detecting human emotions*, 11 2016, pp. 410–415.

[PAR+18] Abhishek Pratap, David Atkins, Brenna Renn, Michael Tanana, Sean Mooney, Joaquin Anguera, and Patricia Arean, *The accuracy of passive phone sensors in predicting daily mood*, Depression and anxiety **36** (2018).

[PTS+16] Niclas Palmius, Athanasios Tsanas, Kate Saunders, Amy Bilderbeck, John Geddes, Guy Goodwin, and Maarten de Vos, *Detecting bipolar depression from geographic location data*, IEEE Transactions on Biomedical Engineering **PP** (2016), 1–1.

[Rus80] James Russell, *A circumplex model of affect*, Journal of Personality and Social Psychology **39** (1980), 1161–1178.

[SA02] Avi Sadeh and Christine Acebo, *The role of actigraphy in sleep medicine*, Sleep medicine reviews **6** (2002), 113–24.

[SC15] Aaron Smith and Pew Reseach Center, *Smartphone use in 2015*, `http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/`, 09 2015.

[SEM91] Peter S. Shenkin, Batu Erman, and Lucy D. Mastrandrea, *Information-theoretical entropy as a measure of sequence variability*, Proteins: Structure, Function, and Bioinformatics **11** (1991), no. 4, 297–313.

[SLS+16] Sohrab Saeb, Emily Lattie, Stephen Schueller, Konrad Kording, and David Mohr, *The relationship between mobile phone location sensor data and depressive symptom severity*, PeerJ **2016** (2016).

[SSH08] Saul Shiffman, Arthur Stone, and Michael Hufford, *Ecological momentary assessment*, Annual review of clinical psychology **4** (2008), 1–32.

[SZK+15] Sohrab Saeb, Mi Zhang, Christopher Karr, Stephen Schueller, Marya Corden, Konrad Kording, and David Mohr, *Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study*, Journal of Medical Internet Research **17** (2015).

[Tek20] Serife Tekin, *Is big data the new stethoscope? perils of digital phenotyping to address mental illness*, Philosophy Technology (2020).

[TKLO16] John Torous, Mathew Kiang, Jeanette Lorme, and Jukka-Pekka Onnela, *New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research*, JMIR Mental Health **3** (2016), e16.

[TRN20] Dirk Tempelaar, Bart Rienties, and Quan Nguyen, *Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application*, PLOS ONE **15** (2020), e0233977.

[TSB+18] J. Torous, P. Staples, Ian J Barnett, Luis R Sandoval, M. Keshavan, and J. Onnela, *Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia*, NPJ Digital Medicine **1** (2018).

[TV20] J. Torous and A. Vaidyam, *Multiple uses of app instead of using multiple apps – a case for rethinking the digital health technology toolbox*, Epidemiology and Psychiatric Sciences **29** (2020).

[TWB+19] John Torous, Hannah Wisniewski, Bruce Bird, Elizabeth Carpenter-Song, Gary David, Eduardo Elejalde, Daniel Fulford, Synthia Guimond, Ryan Hays, Philip Henson, Liza Hoffman, Chun Lim, Michael Menon, Valerie Noel, John Pearson, Randy Peterson, Ammu Susheela, Haley Troy, Aditya Vaidyam, and Matcheri Keshavan, *Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach*, Journal of Technology in Behavioral Science **4** (2019), 1–13.

[VCl+14] Gaetano Valenza, Luca Citi, Antonio lanatà, Enzo Scilingo, and Riccardo Barbieri, *Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics*, Scientific reports **4** (2014), 4998.

[VHT19] Aditya Vaidyam, John Halamka, and John Torous, *Actionable digital phenotyping: a framework for the delivery of just-in-time and longitudinal interventions in clinical healthcare*, mHealth **5** (2019), 25–25.

[WBC+21] Congyu Wu, Amanda Barczyk, R. Craddock, Gabriella Harari, Edison Thomaz, Jason Shumake, Christopher Beevers, Samuel Gosling, and David Schnyer, *Improving prediction of real-time loneliness and companionship type using geosocial features of personal smartphone data*, Smart Health **20** (2021), 100180.

[WPK89] D Williamson, RA Parker, and Juliette Kendrick, *The box plot: A simple visual method to interpret data*, Annals of internal medicine **110** (1989), 916–21.

[WRS+18] Alexi Wright, Nikita Raman, Patrick Staples, Stephanie Schonholz, Angel Cronin, Kenzie Carlson, Nancy Keating, and Jukka-Pekka Onnela, *The hope pilot study: Harnessing patient-reported outcomes and biometric data to enhance cancer care*, JCO Clinical Cancer Informatics **2** (2018), 1–12.

[ZAR+17] John Zulueta, Piscitello Andrea, Mladen Rasic, Rebecca Easter, Pallavi Babu, Scott Langenecker, Melvin McInnis, Olusola Ajilore, Peter Nelson, Kelly Ryan, and Alex Leow, *Predicting mood disturbance severity in bipolar subjects with mobile phone keystroke dynamics and metadata*, Biological Psychiatry **81** (2017), S196.

[Zho09] Zhi-Hua Zhou, *Ensemble learning*, Encyclopedia of Biometrics (2009).

Appendix C

Beiwe Deployment

C.1 Introduction

This appendix contains a detailed description of the deployment of the Beiwe platform. During the deployment, I have encountered several issues, unpleasant obstacles due to the incomplete documentation, and I had to contact the developers several times. It should take a few hours to make a running instance of the Beiwe backend server with these instructions.

C.2 Before Deployment

Prerequisites are

- *Linux* or *MacOS* machine
- *Python 3.6* or newer
- *pip* or *pip3*
- Basic knowledge of *AWS* management console

1. Create an Amazon Web Services *(AWS)* account on *https://aws.amazon.com*. During this process, you have to fill in your credit card number, so it is essential to realise that *AWS* services are not free.
2. Obtain a domain name according to your needs and preferred services.
3. Within *AWS* account set up an *IAM* user with sufficient permissions and download generated credentials:
   - User
   - Access key ID
   - Secret access key
4. Generate deployment key via *AWS* and store it to */.ssh/ directory*, then run:

```
$ chmod 600 path/to/your/key.pem
```

C.3 Deployment

1. To start the deployment, run these commands in terminal to download the repository and check out the development branch. It is essential to work on the development branch as the current master contains bugs and fatal problems.

```
$ git clone https://github.com/onnela-lab/beiwe-backend
$ git checkout development
$ cd backend/cluster-management/general_configuration
$ export DEV_BRANCH ="development"
```

2. Rename *aws_credentials.example.json* to *aws_credentials.json* and fill in stored values from Point 3.

3. Rename *global_configuration.example.json* to *global_configuration.json* and fill in:

```
DEPLOYMENT_KEY_NAME = "Name of the deployment key without
.pem suffix"
DEPLOYMENT_KEY_FILE_PATH = "Absolute file path
with .pem extension"
VPC_ID = "VPC region responding to chosen AWS region"
AWS_REGION = "Location of servers chosen from
AWS management console"
SYSTEM_ADMINISTRATOR_EMAIL = "Email account for error
logging"
```

4. Still in the same directory run:

```
$ pip install -r launch_requirements.txt
$ python launch_script -help-setup-new-environment
When prompted enter preferred environment name
```

5. Edit the following files to turn off Sentry as it changed the architecture and format of their keys, and Beiwe cannot be deployed with the new ones.

```
In beiwe-backend/config/settings.py delete rows (27-29)
corresponding to the following

SENTRY_DATA_PROCESSING_DSN =
getenv("SENTRY_DATA_PROCESSING_DSN")
SENTRY_ELASTIC_BEANSTALK_DSN =
getenv("SENTRY_ELASTIC_BEANSTALK_DSN")
SENTRY_JAVASCRIPT_DSN =
getenv("SENTRY_JAVASCRIPT_DSN")
```

```
In beiwe-backend/cluster_management/deployment_helpers
/configuration_utils.py delete rows 151-154:

for name, dsn in sentry_dsns.items():
    if ensure_nonempty_string(...):
        if not DSN_REGEX.match(dsn):
            errors.append(...)

And rows 159-161:

for key in reference_environment_configuration_keys:
    if key not in beiwe_variables:
        errors.append("{} is missing.".format(key))
```

6. Navigate to the directory *cluster_management/environment_configuration*.

7. In the file *[YOUR-ENV-NAME]_beiwe_environment_variables.json* delete all rows except the one with the domain.

8. The file *[YOUR-ENV-NAME]_server_settings.json* contains default server types for your Beiwe deployment. You do not need to edit this in any way, but you are welcome to do so to minimise the price of the *AWS* services. A study containing fifty people should use the native setup. A quality setup for the study of twenty people is to use the following:

```
"WORKER_SERVER_INSTANCE_TYPE": "t3.medium",
"MANAGER_SERVER_INSTANCE_TYPE": "t3.small",
"ELASTIC_BEANSTALK_INSTANCE_TYPE": "t3.small",
"DB_SERVER_TYPE": "t3.medium"
```

9. Complete following commands:

```
$ cd beiwe-backend/cluster_management
$ python launch_script.py -create-environment
$ python launch_script.py -dev -create-manager
```

10. Install the Elastic Beanstalk Command Line Interface using *AWS* management console.

11. In the beiwe-backend repo, configure the file .elasticbeanstalk/config.yml:

```
branch-defaults:
  development:
    environment: [YOUR-ENV-NAME]
global:
  application_name: beiwe-application
  default_ec2_keyname: [DEPLOYMENT_KEY_NAME]
```

```
default_platform: 64bit Amazon Linux 2018.03 v2.9.4
running Python 3.6
default_region: [AWS_REGION]
profile: eb-cli
sc: git
```

12. In beiwe-backend repository run following commands:

```
$ eb init
$ eb deploy
```

13. In *AWS* management console, find a load balancer of your deployed server and add its *DNS* name to your domain/website using *CNAME* where the target is the *DNS* name.

14. Issue a *SSL* certificate via *AWS* management console.

15. In *AWS* management console, locate to the load balancer's dashboard to see the security section and already assigned source security group. There need to be edited its inbound rules in the following way:

```
HTTP TCP 80 Custom 0.0.0.0/0
SSH TCP 22 Custom 0.0.0.0/0
HTTPS TCP 443 Custom 0.0.0.0/0
```

16. Now you are able to access the dashboard of Beiwe platform using your website.

C.4 After Deployment

1. When you successfully deploy a server, you should be able to login to your site using *https://mysite.example.com.*

2. After login with default admin details *(default_admin, abc123)*, manage studies and create a study.

3. Here you will choose to use production or test study and which data to collect. It is currently essential to choose a test study as the production pipeline does not work, and you can only download raw data.

4. On the main dashboard, add new participants. You can create a different number of participants. Consequently, you will be given details that are supposed to be filled in the app.

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Kubina Michal** Personal ID number: **474555**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Branch of study: **Computer and Information Science**

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Digital Phenotyping Analysis

Bachelor’s thesis title in Czech:

Behaviorální analýza dat

Guidelines:

The goal of this thesis is to get acquainted with digital phenotyping and available mobile platforms which could be used later in medical research and treatment or support of medical patients.
1. Familiarise yourself with the current state of the art of behavioral analysis/digital phenotyping and available mobile platforms which can be currently used for our purposes.
2. Choose and make work the chosen platform and compare it with the others.
3. Make pilot testing and gather data from volunteers.
4. Make an analysis and visualization of the data which were gathered by the volunteers.

Bibliography / sources:

[1] Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder, Gideon, 2016
[2] Recognition of Depression in Bipolar Disorder: Leveraging Cohor and Person-Specific Knowledge, Khorram, 2016
[3] Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech, Zahi N Karam 2014
[4] Realizing the potential of mobile mental health: New methods for new data in psychiatry, John Torous, 2015

Name and workplace of bachelor’s thesis supervisor:

doc. Ing. Daniel Novák, Ph.D., Analysis and Interpretation of Biomedical Data, FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **18.01.2021** Deadline for bachelor thesis submission: **21.05.2021**

Assignment valid until: **30.09.2022**

doc. Ing. Daniel Novák, Ph.D. Supervisor’s signature	prof. Ing. Tomáš Svoboda, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt Student’s signature