Pattern Recognition of Social Contact Events from Wearable Proximity Sensor Data using Principal Component Analysis

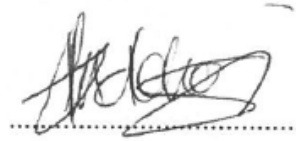
Mvuyo Khuselo Makhasi (460464)

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Engineering.

Johannesburg, June 2019

Declaration

I declare that this dissertation is my own unaided work. It is being submitted to the degree of Master of Science in Engineering to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.



.....

(Signature of Candidate)

19 day of June year 2019

Abstract

Data from wearable proximity sensors can be used to measure and describe social contact patterns between individuals in a household. Previous work describing contact patterns, has been qualitative and relies on visual, subjective observations. Data of this kind has been collected for a short period of measurement ranging from 2-5 days. An automated, quantitative analysis of contact patterns could enable an accurate and new representation of social contact patterns.

Data was collected from ten households, for 21 days in a pilot study implemented in South Africa. 20 datasets were analysed, representing contact events of 20 individuals. Principal Component Analysis was implemented to determine the similarity of contact events across the days of the experiment and to estimate the minimum number of days required to be sampled, to validly represent an individual's contact activity.

The results show that there is a great variation in contact activity across the days of the experiment, as represented by the number of clusters of similar days. The minimum number of days required was determined by the number of days that had a significant contribution to the first three principal components and this varied across individuals from 5 – 11 days. Further analysis on a larger cohort has a potential to provide better social contact parameters for complex social behavioural models and may assist in understanding transmission dynamics of respiratory pathogens, needed in public health research.

In loving memory of my beautiful mother,
Nonzwakazi Ntemi Elizabeth Makhasi

Acknowledgements

I would like to thank first my supervisors, to whom this research would not be possible: Prof Vered Aharonson, Prof Margret Bauer, Prof Cheryl Cohen and Dr Stefano Tempia. The guidance and constructive criticism has shaped this research in a meaningful manner and I will forever be grateful for the contribution you have made. Thank you for your tireless dedication to this research and pushing me to aim for excellence.

Secondly, I would like to thank the Wits Health Consortium (WHC) and The National Institute for Communicable Diseases (NICD), for funding me throughout the course of this research; I truly appreciate the support. I would like to give a special mention to Prof Cheryl Cohen, for supporting me through funding and paying for my Wits residence when I needed the financial assistance the most. Thank you for your kindness and always believing in me.

I would like to thank Dr Ciro Cattuto, Dr Michele Tozzini and Dr Laura Ozella from the Institute for Scientific Innovation (ISI) in Italy, as well as Moses Kiti from the KEMRI Welcome Trust Research Program in Kilifi, Kenya. Their research expertise in the field and collaboration have contributed significantly to this research. I would also like to thank all the research staff from Medical Research Council in Agincourt and Perinatal HIV Research Unit in Klerksdorp. The data used in this research would not be possible, without their assistance. I would also like to thank all the participants in the study, who opened up their homes to me and complied with the study as much as possible, I am indebted to their kindness.

I would like to thank all my close friends, who have been such a pillar of strength and a wealth of inspiration. You know who you are. To my bestie: Welcome Lishivha, I am better off having known you. Thank you so much for all the years, the encouragement and constant reminder for living purposefully. You truly are the friend everyone deserves to have. A special mention to my family (my dad, siblings and precious nephews and nieces) for understanding why I could not always be there, but showed me unwavering support throughout the journey.

Lastly, to the most important and influential person who is no longer with me - my mother. Thank you for the love, the teachings and the beautiful memories. I miss you so much and I hope that this work would have made you proud. This is for you!

Table of Contents

Declaration **1**

Abstract **2**

Acknowledgements **4**

Table of Contents **6**

List of Figures **12**

List of Tables **14**

Glossary of Terms and Abbreviations **15**

Chapter 1 – Introduction **16**

1.1 Background to Problem 16

1.2 Formulation of Research Problem 20

1.3 Research Aims and Significance 21

1.3.1 Research Aims 21

1.3.2 Significance of Research 21

1.4 Structure of Dissertation 21

Chapter 2 – Background **23**

2.1 Ethics Approval 23

2.2 Proximity Sensor 23

2.2.1 Sensing Layer 24

2.2.2 Data Layer 25

2.2.3 Proximity Sensor – Contact Data 25

2.2.4 Proximity Sensor – Status Data 26

2.3 Current Analysis of Proximity Sensor Data 27

2.3.1 Network Analysis 27

2.3.2 Comparison of Proximity Sensor Data and Industrial Alarm Data 28

2.3.3 Pattern Recognition 28

Chapter 3 – Literature Review 31

3.1 Cleaning of Proximity Sensor Data..... 31

3.1.1 Qualitative Assessment on Contact Events 31

3.1.2 Automatic Cleaning of Proximity Sensor Data 32

3.2 Feature Extraction 34

3.2.1 Representation of Contact Events 34

3.2.2 Pattern Recognition 35

3.2.3 Various Applications of PCA..... 36

3.2.4 Application of PCA to Social Contact Events..... 36

Chapter 4 – Experiment Setup 38

4.1 PHIRST Study..... 38

4.2 Logging of Contact Events 39

4.3 Experiment Design..... 39

Chapter 5 – Research Methodology 41

5.1 Mathematical Representation of Proximity Sensor Data 41

5.1.1 Proximity and Status Signals Representation..... 41

5.1.2 Representation of Time Series..... 42

5.1.3 Summary 44

5.2 Cleaning of the Time Series..... 44

5.2.1 Formulation of Cleaning Algorithm..... 44

5.2.2 Summary 45

5.3 Pattern Recognition..... 45

5.3.1 Visual Inspection of Contact Events 45

5.3.2 Principal Component Analysis 46

Chapter 6 – Results 51

6.1 Selection of Data 51

6.2 Data Preparation and Cleaning 52

6.2.1 Data Cleaning Results 53

6.2.2 Percentage Usable Data 55

6.3 Selection of Sampling Period 55

6.4 Presentation of Results 57

6.4.1 Mapping the Variables to the Calendar Days 57

6.5 Class-1: Consecutive Days of Contact Activity 58

6.5.1 Visual Inspection of Contact Events – Time Trends 59

6.5.2 Scree Plot 60

6.5.3 Principal Component Loadings 61

6.5.4 Comparison of Clusters 62

6.5.5 Significance of Days to Principal Components 62

6.5.6 Comparison of Clusters and Significant Days to Calendar Days 63

6.5.7 Summative Comparison of Class-1 Individuals 64

6.6 Class-2: Contact Events for Seven Days or Less 65

6.6.1 Visual Inspection of Contact Events – Time Trends 65

6.6.2 Scree Plot 65

6.6.3 Principal Component Loadings 66

6.6.4 Comparison of Clusters 66

6.6.5 Significance of Days to Principal Components 67

6.6.6 Comparison of Clusters and Significant Days to Calendar Days 67

6.6.7 Summative Comparison of Class-2 Individuals 67

6.7 Class-3: Contact Events for 8-13 Days 68

6.7.1 Visual Inspection of Contact Events 68

6.7.2 Scree Plot 69

6.7.3 Principal Component Loadings 69

6.7.4 Comparison of Clusters 70

6.7.5 Significance of Days to Principal Components 70

6.7.6 Comparison of Clusters and Significant Days to Calendar Days 70

6.7.7 Summative Comparison of Class-3 Individuals 71

6.8 Class-4: Contact Events or 14 Days or More 72

6.8.1 Visual Inspection of Contact Events – Time Trends 72

6.8.2 Scree Plot 73

6.8.3 Principal Component Loadings 73

6.8.4 Comparison of Clusters 74

6.8.5 Significance of Days to Principal Components 74

6.8.6 Comparison of Clusters and Significant Days to Calendar Days 74

6.8.7 Summative Comparison of Class-4 Individuals 75

6.9 Summary 76

Chapter 7 – Discussion 77

7.1 Similarity of Days 77

7.2 Minimum Number of Days Required to Represent Contact Activity 79

7.3 Limitations of Research 81

7.3.1 Quality of Data 81

7.3.2 Representation of True Contact Events 82

7.3.3 Limitations of PCA 83

7.4 Impact of Research 84
7.5 Future Work 84
7.6 Summary 86
Chapter 8 – Conclusion 87
References 90
Appendix A – Ethics Clearance 94
Appendix B – Proximity Pilot Study Design 96
B1. Introduction 96
B2. PHIRST Study 96
B3. Household and Participant Selection 97
B4. Consent Forms and Enrolment 98
B5. Deployment of the Proximity Study 98
B6. Data Quality Control Measures 101
Appendix C – Raw Data Extraction 105
C1. Raw Data Extraction 105
C2. SocioPatterns System Infrastructure 105
C3. Data Extraction 106
Appendix D – Raw and Cleaned Data 110
Appendix E – PCA Results 130
E1. Class 1 130
E1.1 Time Trends 130
E1.2 Scree Plots 131
E1.3 Principal Component Loadings 132
E2. Class 2 133
E2.1 Time Trends 133

E2.2 Scree Plot 135
E2.3 Principal Component Loadings 136
E3. Class 3 138
E3.1 Time Trends 138
E3.2 Scree Plot 140
E3.3 Principal Component Loadings 142
E4. Class 4 144
E4.1 Time Trends 144
E4.2 Scree Plots 147
E4.3 Principal Component Loadings 150

List of Figures

Figure 2.1: Surface View of Proximity Sensor 24

Figure 2.2: Dataframe from Python Notebook Showing the Structure of the Data 25

Figure 2.3: Status Data from Proximity Sensor 26

Figure 4.1: Illustration of Proximity Sensor Environment within a Household obtained from [1]..... 39

Figure 6.1: Raw Data of Individual-1. Each subplot represents a day in the experiment. Y-axis shows the contact events and x-axis shows the time of day.. 54

Figure 6.2: Cleaned Data for Individual 1. Each subplot represents a day in the experiment. Y-axis shows the contact events and x-axis shows the time of day.. 54

Figure 6.3: Percentage Usable Data of Logged Contact Events. The y-axis shows the percentage proportion of cleaned data over the original data. The x-axis shows all the individuals 55

Figure 6.4: Cumulative Variance Representation by the First Three Principal Components for Each Sampling Period. The y-axis shows the percentage cumulative variance of the first three principal components. The x-axis shows the individuals in the study. The cumulative variance is for each sampling period is represented by the various bars on the plot. 56

Figure 6.5: Contact Events Over Time for Individual 11. The subplots i.e. y-axis shows the days of the experiment and the x-axis shows the hour of the day. Each bar represents contact events for that hour. 59

Figure 6.6: Scree Plot Showing Proportion Variance Represented by Each Principal Component 60

Figure 6.7: Loadings Plot for Individual 11. This shows the span of the original variables in the first three principal components space for Individual 11 62

Figure 6.8: Contact Events for Individual 4. The subplots i.e. y-axis shows the days of the experiment and the x-axis shows the hour of the day. Each bar represents contact events for that hour. 65

Figure 6.9: Loadings Plot for Individual 4. This shows the span of the original variables in the first three principal components space for Individual 4. 66

Figure 6.10: Contact Events Over Time for Individual 2. The subplots i.e. y-axis shows the days of the experiment and the x-axis shows the hour of the day. Each bar represents contact events for that hour. 69

Figure 6.11: Loadings Plot for Individual 2. This shows the span of the original variables in the first three principal components space for Individual 2 70

Figure 6.12: Contact Events Over Time for Individual 1. The subplots i.e. y-axis shows the days of the experiment and the x-axis shows the hour of the day. Each bar represents contact events for that hour. 72

Figure 6.13: Loadings Plot for Individual 1. This shows the span of the original variables in the first three principal components space for Individual 1 73

List of Tables

Table 6.1: Summary of Data Used for Analysis 52

Table 6.2: Calendar Grid for 1-to-1 Mapping Between Day Number and Calendar Day 58

Table 6.3: PC Loadings for Individual 1 61

Table 6.4: Summary of Main Results for Class-1 64

Table 6.5: Summary of Main Results for Class-2 67

Table 6.5: Summary of Main Results for Class 3 71

Table 6.7: Summary of Main Results for Class 4 75

Glossary of Terms and Abbreviations

Abbreviation/Term	Description
Dataframe	2-D labelled structured data with columns that are possibly different data types
dB	A unit of measurement, decibel
Edge	Network analysis parameter representing the connection between nodes
EPOCH	Instant of time or date as a point of reference
EVD	Eigenvalue Decomposition
HDSS	Health and socio-Demographic Surveillance Site
ISI	Institute for Scientific Innovation
ISM	Industrial Scientific and Medical
LED	Light Emitting Diode
NICD	National Institute for Communicable Diseases
Node	Network analysis parameter that represents entities that have possible connections to other entities e.g. individuals in a network.
PCA	Principal Component Analysis
PHIRST	Prospective Household observational cohort study of Influenza, Respiratory Syncytial virus and other respiratory community burden and Transmission
PHRU	Peri-natal HIV Research Unit
RFID	Radio-Frequency Identification
Tag	Coin-shaped proximity sensor
Timestamp	A record of the time of occurrence of an event
UID	Unique Identifier
UNIX Time	A system for describing a point in time

Chapter 1 – Introduction

This chapter introduces the research by discussing the background to the problem investigated in section 1.1. This is followed by section 1.2 the formulation of the research problem which includes the research question and its significance. Section 1.3 outlines the research aims and impact. Lastly, section 1.4 outlines the structure of the dissertation.

1.1 Background to Problem

The collection and analysis of digital information of daily activities from mobile devices and social media networks, are changing the understanding of human social interactions and behaviour [1], [2]. In particular, the development of computational models for analysing social contact data, enable necessary predictions of human social behaviour, development of algorithms for advance smart phones and in the medical case, the understanding of epidemiological spread of diseases, through social contact networks [1]–[3]. Contact patterns have been studied in a number of settings that measure age-mixing patterns, distribution of contacts per age group and duration of contacts between individuals. However, the representation and analysis of this type of data, has been largely qualitative and relied on visual and subjective observations. The aim of this research is to apply quantitative methods to represent and identify patterns in contact events between individuals.

Close range physical proximity between individuals has been an area of interest for a number of years [4]–[7]. This type of contact has been important in parameterising models used in agent-based modelling, developing predictive social behaviour models and understanding the contact dynamics between individuals in a given social setting [1]–[4]. Defining contact patterns between individuals includes identifying the individuals in contact, the frequency and duration of these contacts, and where these contacts occur [1], [4]–[7] . There have been a few methods that have been applied to define and quantify these contacts. These include: contact diaries and friendship surveys, Bluetooth devices and smart phones, and wearable

proximity sensors [1], [2], [4]–[8]. All of these methods have merit and some important limitations that are outlined below.

Summarising the work from [5], [6], [9]–[16], contact diaries have been the main methodology for quantifying and describing contact patterns between individuals in a given social setting like a household, a school or work place. Participants are given a questionnaire to identify the individuals they were in contact with, how long those contacts lasted for and how often. Additional information like age and gender is also asked to categorise these types of contacts [5]. This method is subjected to recall bias, under reporting and lack of participation [5], [6]. Other studies like the "Reality Mining Project" in MIT, have implemented Bluetooth devices and smart phones to infer proximity and contact patterns between participants, where Bluetooth logs, call logs and Wi-Fi connections have been used to measure the proximity of participants [4], [8], [17]. These methods have two limitations: the battery life of the devices and the spatial resolution of the Bluetooth logs do not adequately capture close range face-to-face proximity [4], [8], [17]. New technologies have been developed by SocioPatterns [18] that use (radio-frequency identification) RFID technology to measure mutual proximity between tags in contact with one another [1]. This type of technology and the data obtained from it is the focus of this research.

Wearable proximity sensors have been adopted in a number of studies, where participants have been asked to wear a small tag that emits low-power radio frequencies [19]–[24]. All the research papers published on the use of wearable proximity sensors, have calibrated the sensors such that proximity events of <1.5 m are logged [1], [4], [6], [7], [19]–[24]. The devices interact directly with other devices facing each other, in this way it captures close range physical proximity between individuals [1]. There are important limitations in the work that has been done thus far in the analysis of the proximity sensor data. The first limitation is that no automated analysis tools have been developed for identifying contact patterns from the proximity sensor data. The current analysis is based on manual, visually subjective observations and qualitative descriptions of contact patterns.

The typical analysis that is carried out on proximity sensor data, is the construction of contact matrices and contact networks[1], [4], [6], [7], [19]–[24]. A contact matrix is a two dimensional representation of individuals' contacts in a given setting i.e. the rows and columns on a contact matrix represent the individuals in contact and the entries of the matrix are the total number of contacts logged between individuals [1], [4], [6], [7], [19]–[24] These provide the distribution of contacts, the rates and mixing patterns across different age groups [19]–[24]. Contact networks show visually the contact relations within a household [25], [26]. The nodes represent individuals in contact and the links represent the total number of contacts between the individuals [25], [26]. There are several limitations to the current analysis. Firstly, the analysis is qualitative and carries out a summation of all contact events for the duration of the experiment between the individuals. This removes the temporal dynamics that could be observed in the data. The cleaning of the proximity sensor data relies on visual subjective inspection of the contact events i.e. each tag is analysed individually to determine if a contact event is true or not. This becomes a challenge when the number of tags scales up and may not accurately represent the true contact events. A contact network provides visual and some quantitative representation of the contacts, however, the contact events represented as weights or edges in a network, are a cumulative summation of contact events over the entire experiment period. This removes the temporal dynamics, which could provide some important insights into contact patterns.

The purpose of this research is to find a different way of automatically analysing data of this type, by looking at the temporal (i.e. daily and hourly) changes in the contact events and finding patterns that explain and represent the data. The proximity sensor data is event based i.e. a signal only occurs when two or more sensors are in contact (close proximity) [1]. There are two datasets derived from the proximity sensor: the contact dataset and the status dataset [1], [19]. The contact dataset consists of contact event logs which are represented by a power signal in dB between two sensors [1], [19]. These contact events are timestamped and are logged as and when two sensors are in contact with each other [1], [4], [6], [7], [19]–[24]. The status dataset conveys the sensor activity status, which consists of an activity flag which indicates whether a tag is stationary or in motion based on

changes in accelerometer data [1], [19]. These two datasets are sampled at different periods [1], [19]. The contact dataset is sampled at 20-seconds and the status dataset is sampled at 15-minutes [1], [19]. The first objective that will be addressed by this research will be combining the two datasets in a manner that will enable the analysis of both datasets towards the cleaning of the data and identification of contact patterns, without relying on visual and subjective inspection of the data.

Wearable proximity sensors are used as a proxy for close-range physical contact between two or more individuals wearing the sensors [1]. When two individuals wearing the sensors are in close proximity to each other, this is logged as a contact event between the two individuals wearing the sensors [1]. However, when the sensors are not worn by the individuals, but placed in close proximity to each other, the sensors will continue to log contact events [1]. These are false contact events, because they do not represent the individuals, but the sensors in close proximity instead [1]. In this research, the contact and status datasets are used to clean the true contact events. This is done by using the activity status of the sensors i.e. stationary or in-motion to eliminate the false contact events. The assumption that is made in this research is that any contact event that is logged with an activity status value that indicates a stationary tag will be assumed false. Conversely, contact events that are logged together with an activity status value indicating in-motion, will be assumed to be true contact events. Once the true contact events are identified, then the resulting dataset is used to find daily patterns in the contact events

To the knowledge of the author, the only patterns that have been identified from contact events are the contact distribution across different age groups [5], [12]. The contact matrices computed in literature show that children have the highest number of contact events, followed by adults [5], [12]. This is a summation of contact events between the different age groups. Little is known about temporal patterns in the contact events and whether or not these patterns can be clustered into groups. Secondly, studies of this kind that use wearable proximity sensors, have deployed such sensors for 2-5 days and assumed that contact activity between individuals in a household, remains the same throughout the measurement days. This research will determine whether this finding is true and to what extent it can be quantified.

Wearable proximity sensor data is multidimensional which makes it difficult to observe patterns from the raw representation of the data. In particular, temporal changes i.e. hourly or daily changes in contact activity, cannot be easily determined at a first glance. A time series representation of contact events, will be a good start towards finding patterns in the proximity sensor data. Pattern recognition is concerned with the discovery of regularities in input data in order to interpret and represent these regularities [27]. Regularities in data are defined as patterns where there is a repeat or similarity in data trends observed in the data [27]. There are a number of approaches to pattern recognition like statistical methods and machine learning algorithms, which take an input data source and transform it to something meaningful [27]. If patterns exist in the proximity sensor data, by looking at each contact event in time, these can be represented in a manner that can possibly explain and model human contact activity.

In this research, data of logged contact events for individuals enrolled in a pilot study are analysed. Each individual has logged contact events for a 24-hour period over several days of the experiment. This analysis will aim to determine whether there are similarities or differences in contact events across the days of the experiment and what the minimum number of days are required to adequately represent an individual's contact activity.

1.2 Formulation of Research Problem

As indicated in the previous sections, the work done on analysis of proximity sensor data is largely qualitative and based on visual inspection of the data. Very little is known about the temporal patterns in the contact events between individuals. This leads quite nicely to the research question of this research:

Do contact events of individuals demonstrate a pattern of similarity across the days? What are the minimum number of days required to adequately measure and represent an individual's contact activity?

1.3 Research Aims and Significance

There are three main aims for this master's research, which are all outlined in detail below. All of these objectives will attempt to answer the research question mentioned in the above section.

1.3.1 Research Aims

- To develop and apply a method for mathematically representing contact patterns, to enable automated analysis. This will also help towards developing an efficient, automated cleaning algorithm that can be applied to datasets of any size.
- To find similarities and/or dissimilarities in the days of contact activity.
- To determine the minimum number of days required, to correctly represent an individual's contact activity over time.

1.3.2 Significance of Research

This work will contribute significantly towards the implementation and analysis of proximity sensor data.

An automated method for cleaning the proximity sensor data more efficiently will enable proximity sensor datasets of any size to be cleaned in this manner, without relying on visual, subjective observations of the data. Determining the similarity or dissimilarity in days of contact events, will determine the minimum number of days required for representing an individual's contact activity over time. This will lead to better representation of contact patterns, improved parameters in complex social models and assist in modelling transmission dynamics of infectious diseases that rely on contact patterns between individuals.

1.4 Structure of Dissertation

Section 2 provides the background for this research. This gives a detailed context into the nature of the research in terms of the kind of study that is conducted, the infrastructure used and how the raw data is obtained.

Section 3 provides a review of the literature related to the work done and methods that will apply to this research.

Section 4 provides the experiment setup, where and how the research was conducted. This research is bi-disciplinary between the fields of epidemiology and signal processing.

Section 5 details the research methodology used in this research. This includes the representation of the raw data, the cleaning of the raw data and tool used for analysis of the data.

Section 6 presents the results of the analysis on all the contact datasets for each individual in the study.

Section 7 presents the discussion of the results and the arguments that answer the research question. The limitation of the study, as well as recommended future work is also discussed

Section 8 is the conclusion, summarising all the important results and recommendations for potential future work.

Chapter 2 – Background

This chapter provides the background information that contextualises this research. The chapter begins with the ethics statement, which has been approved through a bigger study conducted by the National Institute for Communicable Diseases (NICD). Section 2.2 outlines a detailed explanation of the operation of the proximity sensors, including the type of data that exist. The third section summarises the type of analysis that is and can be applied to this data.

2.1 Ethics Approval

The proximity study, from which this masters research is based on, is nested within the PHIRST study and the protocol has been submitted to the University of the Witwatersrand Human Research Ethics Committee (HREC), which includes the proximity sensor study pilot [28]. The committee has granted the ethics clearance. Participants aged 18 years and older sign written consent forms. For participants younger than 18 years of age, the parent or guardian provides written consent for the child. Assent is obtained from children between 7 years and 17 years by a similar process. Ethics Clearance is attached in Appendix A; clearance number is 150808.

2.2 Proximity Sensor

Wearable proximity sensors have been widely adopted in a number of studies to quantify contact patterns in a number of settings such as hospitals, schools, conferences and rural communities to mention a few [1], [7], [19]–[23]. The aim of these studies was to understand social contact patterns that could possibly drive transmission of infectious diseases [19]. SocioPatterns is a collaboration between three groups of researchers and developers from the Institute for Scientific Interchanges (ISI) in Italy, Centre de Physique Théorique, France and Bitmanufactory, UK which have developed the wearable proximity sensing platform [18]. These devices are active RFID tags which transmit low-power radio-frequency signals between each other [1], [19]. This technology measures direct

face-to-face mutual proximity and represents close interactions between individuals which can be associated to possible transmission of infectious diseases based on direct face-to-face contact [1], [19]. The devices consist of an integrated stack of technologies that enable the monitoring, logging and analysis of human mobility and proximity events [18]. The design of the tags is as shown in Figure 2.1. The proximity tag is a small, light-weight, coin-shaped wearable radio device that individuals can wear embedded inside a badge or a clothing pouch and can sense the close proximity between individuals wearing similar tags [1], [19]. The technology consists of a sensing and a data layer which are summarised from [18], [29] in the following sub-sections.

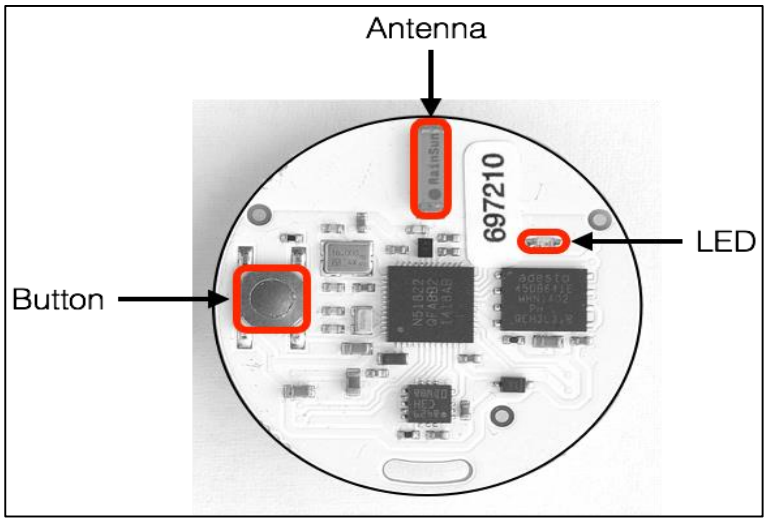


Figure 2.1: Surface View of Proximity Sensor

Figure 2.1, shows the surface view of the proximity sensor. It contains an antenna, an LED and a button that is used to switch to different modes of operation. The tags are embedded with a Nordic Semiconductor nRF51822 microchip and it is deemed safe for use in sensitive environments such as hospitals, schools and households [1]

2.2.1 Sensing Layer

The sensing layer is built on top of the OpenBeacon active RFID devices designed by Bitmanufactory in the UK [29]. The sensing functionality is based on the exchange of ultra-low power radio signals between the wearable devices [1]. Face-to-face proximity of individuals wearing these devices directly facing each other

can be assessed continuously with a temporal resolution of 2 seconds [1]. The devices use the license-free 2.4 GHz Industrial Scientific and Medical (ISM) band of the radio spectrum, which is the same band used by Bluetooth and Wi-Fi networks [1], [18], [29]. This technology is patented in the EU by the ISI Foundation and Bitmanufactory and the cost for each sensor is $25.00 [18]

2.2.2 Data Layer

The SocioPatterns platform comprises both a real-time processing stream and an offline data analysis processing stream [18], [29]. The real-time processing stream is intended to construct live applications and services like real-time visualisation of tags in the sensing environment [18], [29]. The offline data analysis is used for processing recorded data sets and will be the main source for this research, because the focus is not on visualising the tags in the households, but processing the proximity data measured from individuals in the household.

2.2.3 Proximity Sensor – Contact Data

The raw data obtained from the sensors are extracted by using proprietary software developed by SocioPatterns from the ISI Foundation in Italy [18]. As this software is proprietary, the libraries cannot be shared for protection of intellectual property. The NICD and ISI have a collaborative partnership comprising the PHIRST study and the raw data is made available to the researchers. The schema of the raw data looks as shown in Figure 2.2.

t	epoch	uid1	proto	uid2	tx_pwr	rx_pwr	id1	id2
2017-11-04 08:00:19+02:00	1509775219	1841250951	30	1837375701	-16	-89	797249.0	696327.0
2017-11-04 08:00:19+02:00	1509775219	2121794081	30	3626331444	-16	-92	797275.0	797268.0
2017-11-04 08:00:19+02:00	1509775219	4223683453	30	4111321376	-16	-89	797282.0	696459.0
2017-11-04 08:00:19+02:00	1509775219	4223683453	30	3399982893	-16	-89	797282.0	664477.0
2017-11-04 08:00:19+02:00	1509775219	3230102427	30	1841250951	-16	-95	797247.0	797249.0

Figure 2.2: Dataframe from Python Notebook Showing the Structure of the Data

Figure 2.2, shows the Dataframe from python notebook. The “t” column contains the time-stamp. The “epoch” column contains the UNIX time, which can be converted to any time zone for analysis. The “uid1” and “uid2” columns contain the hardware ID’s of tag 1 and tag 2 respectively. The “proto” column contains a flag indicator which indicates a proximity event between the two tags. The “tx_pwr” and “rx_pwr” columns indicate the transmitted and received power packets between the two tags in contact respectively. The last columns “id1” and “id2” are the tag IDs on the sensor itself (meta data). An example for interpreting line 1 in Figure 2.2: at 08:00 on the 04-11-2017, individuals with tags ID *797249* and *696327* were in contact with each other with a power threshold i.e. difference in transmitted (tx_pwr) and received power (rx_pwr) of -75 dB. This data can be analysed using python scripts or exported in a .csv format and analyses using R, Matlab and other related platforms. This data will be analysed in Matlab, because of the powerful analytical toolboxes that come with the package. A detailed explanation of the raw data extraction is provided in Appendix C.

2.2.4 Proximity Sensor – Status Data

Status data is a second dataset obtained from the proximity sensor data. It is logged every 15 minutes, for each sensor, with an associated timestamp.

t	flags	acc_x	acc_y	acc_z	voltage	Household ...	Individual ID
2017-10-09 ...	16	-0.1875	-0.4375	-0.94921875	2.9	K104	K104-004
2017-10-09 ...	0	-0.22265625	-0.45703125	-0.9296875	2.9	K104	K104-004
2017-10-09 ...	16	-0.5625	-0.55078125	-0.70703125	2.9	K104	K104-004
2017-10-09 ...	0	-0.375	-0.41796875	-0.90234375	2.9	K104	K104-004
2017-10-09 ...	0	-0.32421875	-0.30078125	-0.953125	2.9	K104	K104-004

Figure 2.3: Status Data from Proximity Sensor

Figure 2.3, shows a screenshot of the status data. It consists of the timestamp, a *flags* field which is a status code that represents activity state of the tag at the time of the logging e.g. a flag = 0, indicates the tag is stationary and flag = 16, indicates a tag in motion. The changes in accelerometer coordinates are used to determine the value of the flag, which is used to determine whether the tag is stationary or not.

This computation is built into the firmware, developed by ISI. These two datasets (contact and status) are the primary source of analysis for this research

2.3 Current Analysis of Proximity Sensor Data

This section will summarise the methods that can be used to extract features and analyse proximity sensor or comparable data. The section begins with the use of network analysis, a widely used framework for analysing contact events between individuals. Section 2.3.2 summarises how proximity sensors data is analogous to industrial process alarm data and how some of the techniques used in industrial process alarm flood management, can be adapted to proximity sensor data and finally, a section on pattern recognition and how these can be used to find patterns in the proximity sensor data.

2.3.1 Network Analysis

Understanding complex systems, requires knowledge of the interactions of the components that make up that system [25] . A network is essentially a catalogue of a system's components (called nodes or vertices) and the interactions between them referred to as edges or links [25], [26]. Network analysis is widely applied in the analysis of contact events from proximity sensors [25], [26]. Individuals are represented by the nodes of the network and the contact events between the individuals are represented by the links between the nodes [25], [26]. Network parameters like the degree of the nodes, which is a measure of the number of links each node has; average degree of a node which is the probability of a single node having a specific degree; clustering coefficients, which is a measure of probability of neighbours of a node sharing similar nodes [19], [25], [26]. All of these network parameters are used to find patterns that may represent the different types of contact networks. For example, a contact network with a high average degree, indicates a network that it tightly connected where the average node or person is most likely to have a lot of interactions with other individuals in the network [19], [25], [26]. This is important in social contact analysis for public health research in understanding contact patterns that may drive transmission of infectious diseases.

This type of analysis relies on cumulative summation of contact events which may remove some temporal dynamics of contact patterns between individuals. This type of analysis provides important information about contact activity of an individual and age-specific distribution of contact events. Although, network analysis does provide a quantitative representation of contact patterns, it does not provide for the temporal evolution of contact events and it is important to obtain a better estimate of contact events over time.

2.3.2 Comparison of Proximity Sensor Data and Industrial Alarm Data

Proximity sensor data is analogous to industrial process alarm data, where alarm triggers are event-based, with an associated timestamp for each alarm trigger event and the tag IDs of the alarms triggered are also logged [30]–[34]. The proximity sensors data is similarly event-based, which represent contact events that are logged, with an associated timestamp for each event and the tag IDs of the sensors that are in close proximity to each other [1], [4]. The analysis of industrial process alarm flood data can also be adapted and applied to proximity sensor data. An alarm flood is defined as the duration where the rate of alarm annunciation is more than the response capability of an operator [30], [31]. Analysis of alarm flood data is important in root-cause analysis of disturbances in the plant and for alarm flood prediction to warn the operators in time, of an imminent alarm flood [30]–[34]. Similarity measures are used like distance between points and clustering similar points based on the distance score computed [30]–[34]. This technique works well in finding similarities in alarm flood data. Proximity events between individuals in a household could be considered to be similar to process alarm data, where each contact event is similar to an alarm trigger.

2.3.3 Pattern Recognition

Pattern Recognition involves the automatic representation or discovery regularities in a given data [27]. Pattern recognition has its origins in the engineering discipline, whereas machine learning grew from the field of computer science [27]. There are different approaches to finding patterns in data which include, but are not limited to: taking an input data and assign a label or class to it; clustering groups of input data, based on similarities in the features of the data and making predictions about

future outcomes based on statistical analysis [27]. The paragraphs below outline some of the known methods used in pattern recognition and machine learning.

Supervised Learning: Supervised learning involves the use of classification algorithms that take a set of input data, referred to as a training set, to assign a label or class to each input [27]. These algorithms are used when the labels are known [27]. Supervised algorithms are used in classification problems, regression models and play an important role in pattern recognition [27]. The patterns in proximity sensor data are unknown and it is difficult to assign labels to contact event data, when clusters (if they exist) are not known or well defined. A supervised learning algorithm, will work well once clusters in the data are known with the associated labels.

Unsupervised Learning: these define the type of problems where the labels or classes are not known and machine algorithms are used to identify groupings or similarities in the data [27]. These type of algorithms are important when pattern in a given data set are not known and as such, data points with similar attributes are clustered together [27]. Unsupervised algorithms play an important role in generating hypotheses or new knowledge from the data [27]. This type of algorithms could be applied towards pattern recognition of contact event data [27].

Dimensional Reduction: Some data sets are multidimensional and often, patterns are not easily observed from the original data [27]. Such data sets need to be transformed, by reducing the number of dimensions in order to observe patterns in a lower dimensional space [35]. Dimensional reduction is also important in selecting the best features that represent or explain the most variation in the data [35]. Algorithms that are widely used for dimensional reduction, include Principal Component Analysis, which is a statistical method for reducing the number of features required to represent data and do exploratory analysis of patterns in the data [35].

PCA is a chosen method for the analysis of the proximity sensor data in this research. The proximity sensor data is multidimensional and to the knowledge of the author, the temporal features of contact event data have not been analysed to explore the contact patterns that may not be discovered by the current methods used in literature. This will provide principal components that will account for the most variation in the data and these will be used to find patterns that will represent contact events between individuals wearing the sensors.

To summarize, proximity sensors act as a proxy for close-range face-to-face communication between individuals in a given setting. The proximity sensors log contact events as they occur. However, as outlined in the previous section, this data is noisy and up to this point, there has been no automated analysis carried out on contact events from proximity sensor data. The next section will document the literature review related to the work done on the analysis of the contact events and methods that will are adapted for this research.

Chapter 3 – Literature Review

This chapter reviews the literature that relates to the work that is presented in this research. The first section discusses the current methods used for pre-processing proximity sensor data and their limitations that will be addressed by this research. The second section describes the feature extraction using Principal Component Analysis technique and pattern recognition.

3.1 Cleaning of Proximity Sensor Data

As mentioned in the introductory chapter, proximity sensor data contains a number of misrepresentations of true contact events. These are false contact events that only represent the proximity event between sensors, when individuals are not wearing the sensors. In this section, the methods that have been applied in literature towards the cleaning of proximity sensor data, will be reviewed. Specifically, this section will review the work done by *Kiti et al* and *Ozella et al* [19], [36], whom have both used the contact and status datasets to clean out false contacts contact events.

3.1.1 Qualitative Assessment on Contact Events

The contact dataset consists of timestamped contact events, sampled at 20 seconds. The cleaning approach followed by *Kiti et al* and *Ozella et al* [19], [36], is to aggregate the contact events by hour and plot these for a chosen period of time. These types of plots will show the temporal evolution of the contact events. The assumption they make is that if there is no change in contact activity for an extended period, then those logged contact events are most likely to be false contact events, which could imply that the sensors were placed together and not worn by the participants [19], [36]. This form of analysis is repeated for each individual sensor in the sample [19], [36]. From the status dataset, the flag field is used to determine the activity state of the sensors. The status values are logged in memory every 15 minutes as a flag value of either 0 and 16, which relates to stationary and in-motion state, respectively [19], [36]. There are other flag values, which are related to

memory error, memory size, battery level etc. that are not of interest in this particular case.

Using the flag status, *Kiti et al* and *Ozella et al* [19], [36], plotted the flag status values i.e. 0 and 16 over time to identify periods of inactivity, where the tag was not active or in motion. In their work, they assumed the periods of inactivity would be related to the evening time where the participants are expected to be asleep and not wearing the sensors [19], [36]. As a result of this, *Kiti et al* and *Ozella et al* were only interested in cutting out these periods of inactivity i.e. evening times where the sensors were not worn [19], [36]. In their work, they define a time of active period i.e. day-time as the period between 06:00 and 22:00 and the period of inactivity i.e. evening between 22:00 and 06:00, where any contact events logged between the period of inactivity would be cut out from the dataset as noise [19], [36].

Although this plays an important role in cleaning out unwanted contact events as in the case of the evening, it still assumes that during the day or active periods, there would not be contacts that are false. The flag values will be used in this research to make a distinction between contact events and proximity events. The assumption that will be made is that if a flag value of 16 is recorded, then any proximity event that is logged for the same time point should be a true contact event, because a tag in motion will represent an individual wearing the tag. In order to achieve this, the two datasets will have to be synchronised i.e. sampled at the same time. A series of event-based status data will be transformed to a discrete signal with equally spaced samples. The next section will show the methods that can be applied to synchronise the dataset and ultimately clean the contact events.

3.1.2 Automatic Cleaning of Proximity Sensor Data

In this section, the methods considered for the automatic cleaning of the proximity sensor data are reviewed. This includes the time series representation of contact events, representation of event based data and the actual proposed method for cleaning.

Time Series Representation of Proximity Sensor Data

Proximity Sensor Data is analogous to industrial process alarm data as explained in section 2.3.2 [30]–[34]. Both Industrial Process alarm data and proximity sensor data are event-based, multidimensional datasets that represent an alarm trigger and proximity events respectively [1], [4], [30]–[34]. This comparison is important, because the representation of process alarm data can be applied to proximity sensor data. For instance, in alarm flood management an alarm count for each period is calculated, which represents the total number of alarms that are triggered in a given period [30]–[34]. Similarly with proximity sensor data, in a given period of 20-seconds, a single sensor can log up to three proximity events with two or more sensors in the same proximity sensing environment [1], [4], [19], [36]. Proximity sensor data can be represented as time series data of proximity event counts, adapted from alarm count used in industrial processes. This is important for cleaning of the data by having a time series representation of proximity events and the activity status of the sensors.

Event-based Versus Equidistant Sampling

Signal processing allows for the representation, transformation and manipulation of signals and the information contained by them [37], [38]. The proximity sensor data consists of a series of event based signals, with two data sets (contact and status) that are sampled at different periods [1], [4], [19]. The contact data is sampled every 20 seconds and the contacts are logged as they occur, whereas the status data is logged every 15 minutes, which makes the two data sets asynchronous [1], [4], [19]. In order to overlap the contact proximity sensor data with the status data, these data sets will be transformed and sampled at the same sampling period, in order to generate a series of discrete signals, with equally spaced samples [37]–[39]. The missing values of status dataset will be populated by holding the previous known value to the next known value. In this way, the contact dataset and the status datasets can be aligned on the same time grid. Matrix representation of the contact dataset and the status dataset is done in order to carry out an elementwise multiplication of the two matrices. This will determine the true contact events from the false contact, explained below.

Data Cleaning

The proximity sensors data provides two information i.e. the logged contact events and activity status of the proximity sensors. Using matrix representation of the two information, an element-wise multiplication of the matrices, called a *Hadamard* product [40] [41], can be used to eliminate contact events where the activity status of the sensor is stationary. The assumption is that if the sensors is in stationary state, then that particular sensor is not worn by the individual. Therefore, any contact events logged by the sensor in stationary state can be assumed to be false.

3.2 Feature Extraction

This section will describe how the features for the contact events will be extracted and how these will be used towards finding patterns in the data. The first part of this section will look at the mathematical representation of the data, followed by the application of Principal Component Analysis, which is the proposed method for analysis of contact events in this research.

3.2.1 Representation of Contact Events

Cattuto et al [1], developed a framework for the representation of contact events from proximity sensors. This framework defines a contact event, as a single event that lasts for at least 20-seconds between any pair of proximity sensors. This framework is used in this research to represent each contact event between two individuals as a time series of events, with equally spaced samples of 20-seconds. Using, the concept of industrial alarm count, each contact event that has been logged in the same 20-second period will be summed as a contact event count. This will yield a time series representation of contact events. A contact event matrix, from this time series representation of contact events, is computed. This contact event matrix, is used to find hourly and daily patterns of contact events between individuals.

The literature pertaining to studies that have implemented wearable proximity sensor data, shows that studies of this kind have measured contact events for a short period of time. Specifically, the number of days used to measure contact events

between individuals in a given setting ranges from 2-5 days and the number of days used for analysis are typically 2-3 days of measurement [7], [21], [23]. A cosine similarity has been implemented by Kiti et al and Ozella et al to measure the similarity of contact events over time [19], [36]. In their work, they have found that individual's contact activity remains fairly similar over time and that a few days of measurement should adequately represent an individual's contact activity. Through pattern recognition, a number of minimum days required for measurement will be investigated in this research.

3.2.2 Pattern Recognition

Principal Component Analysis is a simple and easy method to implement that allows for dimensional reduction of highly correlated datasets, feature selections and pattern recognition in the reduced dataset [42]–[45]. Visualisation of patterns in a highly correlated dataset does not reveal easily the dynamics of the dataset [42]–[45]. PCA allows for this dimensional reduction, without losing much the variation in the original data [42]–[45]. In this way, the features that best represent the original data can be selected and patterns can be explored in the resulting dataset, which reveal the hidden dynamics that could otherwise not be observed in a higher dimension [42]–[45]. This method is used widely in signal processing, fault diagnosis in industrial processes and pattern recognition in a number of fields such as image processing, event-based data analysis and industrial processes [46]–[49].

PCA transforms a set of highly correlated variables to a new set of orthogonal, uncorrelated variables called principal components [50]. This transformation is important to extract the most important information from the original variable, by retaining the original variance in the data [50]. In this way, principal components that represent the least amount of variance can be discarded such that the dimension of the original data can be reduced, without losing much variance [50], [51]. Patterns can be identified in the reduced principal component dimensional space through exploratory analysis [50].

3.2.3 Various Applications of PCA

PCA is widely used technique for dimensional reduction, by reducing the multiplicity of a highly complicated dataset, while retaining the variation in the original data [42]–[45]. In Wireless Sensor Networks, PCA is implemented in the aggregation process of sensor nodes, by reducing the number of packets transmitted and battery consumption at each sensor node [42]. In image processing, PCA is applied in object identification in an image, by selecting the best features that account for the most variation in the data [46]–[49]. This technique is applied facial detection algorithms, object identification in computer vision and in image compression algorithms [46]–[49].

In industrial processes, robots are used to automate repetitive tasks and thus increase productivity [39], [43]. Often faults occur and sometimes the delayed response to various faults in plants may have cost implications [32]–[36] [39], [43]. Fault prediction models are used to optimise the detection of faults before they occur [39], [43]. PCA has been applied in fault detection models that use event log data both written and logs from robot controllers, to reduce the dimension from log event data to detect abnormal behaviour from event log patterns [43].

Root-cause analysis is often carried out in industrial process to analyse the propagation of faults through an industrial production process [52]. This is done by investigating measurements of process variables like pressure and flow, but this is often limited to one section of the plant [52]. PCA enables the multidimensional representation of process measurement variables, for each section of a plant and determine which measurements contributes the most to the observed disturbances in the plant [52].

3.2.4 Application of PCA to Social Contact Events

PCA represents a dataset of multiple inter-correlated variables, by a set of new uncorrelated variables called principal components [50]. The goal of PCA is to extract the most important information in a dataset, represented by the variation in the data and compress the size of the data by retaining only the components that contain the most important information [50]. Variables that do not contain

important information are redundant and can be discarded from the analysis [50], [51]. Applying this method to social contact events, the similarities in contact events across the days can be determined. Using feature selection, the variables or days that have a significant contribution to the principal components can be determined and used to determine the minimum number of days required to describe an individual's contact activity [51], [53].

In summary, extensive previous research sought to describe contact patterns between individuals using wearable proximity sensor data. All of the work done up to this point has relied on qualitative assessment, with visual observations of contact patterns and some quantitative representation of contact events through network analysis. The only contact patterns that are known from data of this kind is the age distribution of contact events, which show that children have the highest number of logged contact events. This method carries a cumulative summation of contact events over the experiment period. This removes the dynamics in the contact event data over time. PCA was successfully used for many other applications and signals, which has similar properties to the proximity sensors data. This method could represent the temporal patterns of contact events and extract the most important features that represent the most important information in the contact data.

Chapter 4 – Experiment Setup

In this section, a summary of the study design is presented. A pilot study was designed and deployed by the author, for a period of three weeks in two sites, namely: Agincourt in Mpumalanga and Jouberton in Northwest. A summary of the experimental design is provided in this section. Section 4.1 provides the context in which this experiment was designed and conducted. Section 4.2 describes the data collection through the logging of contact events from the wearable proximity sensors. Section 4.3 outlines the design of the experiment.

4.1 PHIRST Study

This research is nested in the PHIRST Study (Prospective Household observational cohort study of Influenza, Respiratory Syncytial virus and other respiratory pathogens community burden and Transmission dynamics in South Africa) [28]. The PHIRST Study is implemented by the National Institute for Communicable Diseases (NICD) and is a household-level community cohort study implemented in two sites in South Africa [28]. The first site is a rural community in the Mpumalanga Province, where the study is conducted within the Health and socio-Demographic Surveillance Site (HDSS) at the Medical Research Council (MRC)/ Wits Rural Public Health and Health Transition Unit (in Agincourt).

The second site is a semi-urban community in the Jourberton Township in Klerksdorp, in the North West Province, under the Perinatal HIV Research Unit (PHRU) of the Chris Baragwanath Hospital. This study ran for a period of three years, from 2016-2018. The data for this research is obtained from a pilot study ran by the author within the PHIRST study in 2017, where ten households were approached and enrolled and individuals were asked to wear the proximity sensors for three weeks. The aim is to use these datasets to automatically identify contact patterns across the days, between individuals at a household level.

4.2 Logging of Contact Events

The proximity sensor is a radio frequency wearable sensor used to detect mutual proximity between individuals by exchanging power packets that are translated into a proximity measurement [1].

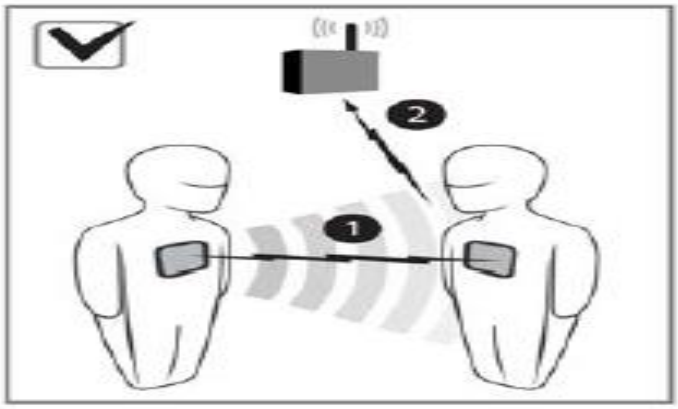


Figure 4.1: Illustration of Proximity Sensor Environment within a Household obtained from [1].

Figure 4.1, shows an illustration of a proximity sensor environment obtained from [1]. Individuals wear a proximity sensor in a lanyard, placed in front of their chest as shown in Figure 4.1. The radio frequency power packets, shown by number 1, are logged as contact events and these represent the close range physical proximity between individuals, within a distance of <1.5 m. Number 2 in the illustration, shows a reader that can be used to visualize the proximity events between sensors, but this technology was not used towards this research

4.3 Experiment Design

Ten households were selected randomly, five households in both the rural and urban site. All the participants in the study were asked to wear a small, coin-shaped proximity tag to measure mutual proximity between participants in contact with each other. The participants were required to wear the tags throughout the day and in the evening. Routine visits were conducted by the field workers and the author, to check the operational state of the tag, ensure that the tag itself was worn correctly

and to assist with any concerns that the participants had. A detailed outline of the experiment design and operational procedures is included in Appendix B.

In summary, a pilot study was conducted by the author in two sites in South Africa: a rural and a peri-urban community. Wearable proximity sensors were deployed to participants who have consented to participate in the study. Data from these sensors was used to quantity the contact activity between individuals wearing the sensors. It is this data that is used to determine the similarity of days of contact activity between individuals in a household and to determine what the minimum number of days of measurement, to adequately represent the contact activity of an individual. The next chapter will present the methodology used for analysis of data of this kind.

Chapter 5 – Research Methodology

This chapter explains the research methods outlined for this master's research. The data that will be used for analysis comes from a pilot study that has been conducted by the author in October 2017. The first section in this chapter, begins with a mathematical representation of the data. Section 5.2 details the cleaning algorithm. Section 5.3 presents the application of the PCA method for pattern recognition.

5.1 Mathematical Representation of Proximity Sensor Data

In this section, the mathematical representation of proximity sensor data is provided. The mathematical formulation derived in this section is used towards the quantitative analysis of contact events from the proximity sensor data.

5.1.1 Proximity and Status Signals Representation

Proximity sensor signals are event-based signals that only exist at specific points in time. A proximity event: $P(X_i, X_j, t) \in$ [-16; -70], is a logged event between two sensors X_i and X_j at time *t*, measured in dB.

At time *t*, a proximity event will occur if and only if, two sensors come in close proximity to each other, otherwise, no proximity event will be logged i.e. $P(X_i, X_j, t) \neq 0$ iff X_i and X_j are in contact. This contact event can be further defined as:

$$C(X_i, X_j, t) = \begin{cases} 1, & if\ P(X_i, X_j, t) \neq 0 \\ 0, & when\ no\ contact\ occured \end{cases} \quad (1)$$

Where the value of 1 and 0 represent a logged proximity event and no proximity event, respectively. An analogy can be made to industrial process alarm count, where there can be more than one alarm events logged within the same period. In this study's data, a single sensor can log up to three proximity events between other sensors for the same time period in 20-seconds. The analysis will focus on the total number of proximity events logged by each sensor at time *t*. This is given by:

$$C(X_i, t) = \sum_{j \neq i} C\left(X_i, X_j, t\right) \tag{2}$$

Where *i* is the selected individual in the household whose contact event count is calculated. For example, consider an individual *i=1*, from household of *4* individuals, the contact event count at time *t* is given by:

$$C(X_1, t) = \sum_{j \neq i}^{4} C\left(X_1, X_j, t\right)$$
$$= C(X_1, X_2, t) + C(X_1, X_3, t) + C(X_1, X_4, t)$$

The contact event count will give the contact activity of each sensor in terms of total number of proximity events at each time interval. This representation of proximity events is important when comparing the proximity dataset with the status dataset (described below), towards the cleaning of contact events from proximity events.

A status flag is represented by:

$$S(X_i, t) = \begin{cases} 16, & if\ sensor\ X_i\ \ is\ in\ motion \\ 0, & if\ sensor\ X_i\ is\ \ stationary \end{cases} \tag{3}$$

Where X_i, represents the individual sensor and *t* is the time point at which the flag field was logged.

5.1.2 Representation of Time Series

Each sensor data can be represented as a vector of proximity events for each 20-second interval for 24 hours, for each day of the experiment. Let $C_i(d, k)$ denote the contact event vector for individual *i*, where *d* represents the day number e.g. *d = [1; m]*, where *m* = last day of experiment and *k* is the discrete sample number. The contact event vector for a single day of contact events is given by:

$$C_i(d,k) = [\, C_1(1,1)\;\; C_1(1,2)\;\cdots\;\; C_1(1,k)\,] \tag{4}$$

Where *i* represents the individual number, *d* represents the day and is equal to 1 in this case; *k* represents the discrete sample i.e. *k = [1; N]* where *N* is the total number of samples for each day = 4320. A daily contact event matrix can be defined using these vector representations, for each individual sensor, given by

$$C_i(d,k) = \begin{bmatrix} C_i(1,1) & \cdots & C_i(1,k) \\ \vdots & \ddots & \vdots \\ C_i(d,1) & \cdots & C_i(d,k) \end{bmatrix} \tag{5}$$

Where *i* represents the individual, *d* represents the day number and *k* is the sample. Each row in the matrix represents contact events for a given day e.g. day = 1, from sample k = 1 to sample N = 4320. The last row, represents contact events for the last day = m of the experiment. Similarly, with the status dataset, a daily status matrix can be defined as follows

$$S_i(d,k) = \begin{bmatrix} S_i(1,1) & \cdots & S_i(1,N) \\ \vdots & \ddots & \vdots \\ S_i(m,1) & \cdots & S_i(m,N) \end{bmatrix} \tag{6}$$

The matrices in (5) and (6) will be used to clean the dataset, by comparing each matrix element in the daily contact matrix to that of the daily status matrix.

5.1.3 Summary

Re-capping from the previous section 5.1, true contact events are assumed to have occurred when a logged contact event is matched with an active status flag i.e. the tag is in motion with $S_i(d,k)$ = 16. This way, true contacts can be extracted from the rest of the proximity dataset.

The following attributes need to be cleaned from the proximity sensor data

1. Proximity events between sensors only.
2. Contacts that have been logged, but with a status flag of 0, indicating the tag is stationary or not in use.

5.2 Cleaning of the Time Series

In this section, the *Hadamard* product will be computed between the daily contact matrix in (5) and the daily status matrix in (6), to yield a matrix of true contact events between individuals wearing the sensors.

5.2.1 Formulation of Cleaning Algorithm

Given a daily contact matrix **C** and daily status matrix **S**, both having the same dimension *M-by-N*, where *M* represents the days in the experiment and *N*, the contact events over a 24-hour period, sampled at 20 second intervals, yielding 4320 samples. Define a daily contact event matrix **E** as:

$$\boldsymbol{E} = \boldsymbol{C} \circ \boldsymbol{S} \tag{7}$$

Where each element of **E** is

$$\boldsymbol{e}_{ij} = (\boldsymbol{C})_{ij}(\boldsymbol{S})_{ij}$$

The resulting matrix **E**, is the daily contact event matrix of dimension *m* x *n*, representing the true contact events between individuals wearing the sensors. This matrix is given by

$$E_i(d,k) = \begin{bmatrix} e_i(1,1) & \cdots & e_i(1,N) \\ \vdots & \ddots & \vdots \\ e_i(m,1) & \cdots & e_i(m,N) \end{bmatrix} \quad (8)$$

Where *i* represents the individual, *d* is the day of the experiment from day 1 to *m* (last day) and *k* is the discrete sample number from $k = 1$ to N, where $N = 4320$.

5.2.2 Summary

Up to this point, no analysis has been carried out on identifying contact patterns in the data. The first objective was to clean the data by identifying what are true contacts events and removing false contacts that only represent the sensors in proximity. The next objective is to find patterns in the cleaned data and analyse these. Principal Component Analysis is used to represent the original variation in the data in a reduced dimensional space; explore patterns in the data and find clusters in the data based on similar attributes.

5.3 Pattern Recognition

In this section, the methods for finding patterns in the contact events are presented. This section begins with the visual inspection of the contact events and finding days of similar contact activity. The second section, presents the application of Principal Component Analysis in reproducing the patterns observed by visual inspection, in a quantitative manner.

5.3.1 Visual Inspection of Contact Events

A plot of all contact events over the number of days of the experiment will be done for all individuals in the study. There will be as many subplots as there are days in the experiment and each subplot represents the contact events over the 24-hour period in day. Days with similar periods of contact activity will be identified through visual inspection and clustered together. Days that exhibit differences in

contact activity will be identified as outliers. The measure of similarity through visual inspection is

1. Days where contact activity is logged at similar times in the days

2. Period of similar contact activity during the day i.e. similarity in morning, afternoon and evening periods.

Days that show similarity based on the above criteria for visual inspection, will form a cluster of similar days. The days that show variation in contact activity based on the above criteria will not form a cluster. Visual inspection of the data alone, will not be sufficient to represent the patterns in the contact activity. Principal Component Analysis provide a quantitative representation of the structure and relationship between the observations and variables in the dataset [50]. The clusters observed through visual inspection will be compared to those identified through principal component analysis.

5.3.2 Principal Component Analysis

Principal Component Analysis is used in this research to determine the variation in contact activity and find clusters of similar days. The purpose of Principal Component Analysis is to quantify and summarise the important information in a dataset that is described by multiple inter-correlated variables [50]. PCA then extracts the most important information in a dataset, represented by the variation that exists in the data, and represents this as a set of new, uncorrelated variables called principal components [50], [54]. These principal components are linear combinations of the original variables. In so doing, the variables that contribute the most to each principal component, can be considered to carry the most information in the data and those that do not, are considered to add redundancy to the data and can be removed [54].

Overview of Principal Component Analysis

Let $\boldsymbol{X}$ denote a data matrix with M-by-N dimensions, where M is the number of observations and N is the number of measurements or features. PCA uses eigenvalue decomposition of a covariance matrix $\boldsymbol{C}$ given by:

$$\boldsymbol{C} = \boldsymbol{X^T X} \tag{9}$$

Where $\boldsymbol{C}$ has dimensions *N-by-N*, where the number of dimensions of $\boldsymbol{C}$ is equal to the number of variables or columns of data matrix $\boldsymbol{X}$. The covariance matrix in (1) is decomposed into a set of eigenvectors $\boldsymbol{W}$ using eigenvalue decomposition, given by

$$\boldsymbol{W^T} = [\boldsymbol{w_1} \ \ldots \ \boldsymbol{w_M}] \tag{10}$$

Where $\boldsymbol{W}$ is a *N-by-N* matrix, of *N* orthogonal basis vectors called principal components. The principal components are a set of new variables, which are a linear combination of the original variables of data matrix $\boldsymbol{X}$. The columns of $\boldsymbol{W}$ represent the principal components and the rows represent the loadings, which are the contribution of the original variables towards each principal component. The eigenvalue decomposition of the covariance matrix $\boldsymbol{C}$ will also yield *M* eigenvalues d_m. The linear transformation of the observations of $\boldsymbol{X}$ is given by

$$\boldsymbol{T} = \boldsymbol{W^T X} \tag{11}$$

Where matrix $\boldsymbol{T}$ in (11) is called a score matrix and has dimensions *M*-by-*N*. The columns of matrix $\boldsymbol{W}$ are ordered by the highest value of d_m in decreasing order. The first *R* components that contribute the most to the variation of the original data are selected, such that $R < N$. In this way, the dimensions of the score matrix are reduced, based on the value of R selection. The new score matrix is given by

$$\boldsymbol{T}_R = \boldsymbol{W}_R^T \boldsymbol{X} \tag{12}$$

Where the dimensions of matrix $\boldsymbol{T}_R$ are *M-by-R*. The value of *R* can be selected based on the number of principal components that have the most contribution to the variation of the data. A loadings and score plot can be made using the matrix in (10) and (12), respectively. This will show a plot in the R-dimension space and patterns can be observed from these plots. External known information about the data like meta data or any process specific information, depending on the type of data used, can be applied to cluster the points, this is called exploratory analysis where hypotheses can be generated based on the patterns that can be learned from the data

Application of PCA to Wearable Proximity Sensor Data

The PCA method outlined in section 3 is applied to visualise the patterns of contact events for each individuals in the experiment. The variables are the days of the experiment and the observations are the contact events over the hours of the day. The sampling period that provides the best representation of the variance, will be determined. The goal is to observe if there are similarities or differences in day of contact activity for each individual in the experiment.

Let input data matrix **H** be given by

$$\mathbf{H} = \boldsymbol{E^T} \tag{13}$$

Where

$$\boldsymbol{H}(i,j) = \begin{bmatrix} h(1,1) & \cdots & h(1,N) \\ \vdots & \ddots & \vdots \\ h(m,1) & \cdots & h(m,N) \end{bmatrix}$$

Where matrix **E** is obtained from (8) and data matrix **H** in (13), represents the hourly contact events of an individual in the experiment, *i* the sample number i.e. contact events from the first to the last hour of the day i.e. *i = [1; m]*, where *m* = 24; *j* represents the day of the experiment i.e. *j = [1; N]* where *N* = last day of the experiment. In simple terms, each row in the data matrix **H** represents the contact events over the hours and each column represents the days of the experiment. **H** is a *M-by-N* matrix.

Principal Component Analysis using Eigenvalue Decomposition

Given matrix **H** as the input matrix for principal component analysis, the covariance matrix of **H** is given by

$$\boldsymbol{C_H} = \boldsymbol{H^T H} \tag{14}$$

PCA computes the loadings matrix of (14) using Eigenvalue Decomposition and is given by:

$$\boldsymbol{W^T} = [\boldsymbol{w_1} \ \ldots \ \boldsymbol{w_N}] \tag{15}$$

The loadings matrix in (15) is an *N-by-N* matrix of principal components and are ordered by highest value of the eigen value d_m.

Selection of Components

A scree plot shows the proportion representation of variance by each principal component [51]. This plot conveys how much information or variance is retained by each principal component. Using this plot, the total number of principal components that represents a desired level of variance are chosen [50], [51]. The first *R* components will be selected such that $R < N$. This can be done using the elbow method or the desired variance represented of the first *R* components. The resulting score matrix is given by

$$\boldsymbol{T}_R = \boldsymbol{W}_R^T \boldsymbol{H} \tag{16}$$

The score matrix in (16) has *M-by-R* dimensions. The score matrix provides the projection of the original variables onto the principal component space. The loadings matrix provides the contribution of each original variables to the principal components. The analysis in this research will focus on the loadings plot and contribution of the original variables to the first *R* principal components

Clustering of Days from Loadings Plot

The loadings plot represents the original variables in the first *R* principle components using the loading values as coordinates in the R-D principal component space, provided that $R \leq 3$ [50]. Variables that are similar will span in the same direction of the principal components [50]. The closer the variables are in a quadrant, the more similar they are to each other. The variables that are different from each other, will span in the opposite direction or be further apart. Variables that are similar will form a cluster, in these cases days that span in the same direction and close to each other, will form a cluster. These clusters are compared to those observed through visual inspection of contact events.

Significance of Variables

As explained earlier, Principal Component Analysis transforms a set of inter-correlated variables to a set of new uncorrelated variables called principal components [54]. Each principal component is a linear combination of the original variable [50]. Variables that have a large variation i.e. carry a lot of information in the data will have the strongest contribution to the first few principal components. They represent how strongly correlated each original variable is to each principal component. Each principal component can be represented mathematically using (17):

$$PC = \sum_{i=1}^{N} \alpha_i \, x_i \tag{17}$$

Where $\alpha = [\alpha_1 \dots \alpha_N]$, representing the total contribution of each variable to the first R principal components and $x_i = [x_1 \dots x_N]^T$, representing the days of the experiment. From (17), the absolute value α determines the contribution of each variable. The smaller the value of α, the less significant the associated variable is. When the contribution of a variable is insignificant in the principal component space, it can be assumed that the variable is also not important in the original measurement space [50], [54]. A threshold can be set to discard variables that are not significant and in this research, the total number of variables that are significant are assumed to represent the minimum number of days required to represent the contact activity of an individual.

Chapter 6 – Results

This chapter presents the results from the analysis of the datasets obtained from the pilot study conducted by the author. The aim of this chapter is to present findings that will answer the research question in this master's research: how similar are individual's contact events across the days of the experiment and to determine the minimum number of days required to represent an individual's contact activity. Section 6.1 begins with the selection of the data used in this research, followed by section 6.2, which outlines the data pre-processing and the cleaning results. Section 6.3 describes the sampling periods considered and the final chosen sampling period for analysis. Section 6.4 outlines how the results are presented. Section 6.5-6.8 presents the results for the individual datasets. Section 6.9 provides a summary of the results.

6.1 Selection of Data

A pilot study was conducted in two sites in South Africa as outlined in Chapter 4. Data from 45 individuals from ten households was collected. Each participant was asked to wear a proximity sensor which logged contact events between individuals who were in close range proximity of <1.5m, for a period of three weeks. Every individual's contact events are represented as an independent dataset. Out of the 45 datasets, only 20 datasets had usable data and this was determined by a minimum of five days or more, of logged contact activity. The results of the analysis of the data collected from these sensors are presented in this chapter.

Table 6.1: Summary of Data Used for Analysis

Individual ID	Study ID	Age	Gender	Available Time Trends	Number of Days for Analysis
Individual 1	A112-001	44	F	2017-10-13 to 2017-10-30	16
Individual 2	A112-003	22	F	2017-10-13 to 2017-10-30	13
Individual 3	A112-004	7	M	2017-10-13 to 2017-10-30	9
Individual 4	A112-006	4	F	2017-10-18 to 2017-10-30	7
Individual 5	A112-007	20	M	2017-10-13 to 2017-10-30	15
Individual 6	A125-001	64	F	2017-10-13 to 2017-10-29	15
Individual 7	A125-004	14	F	2017-10-13 to 2017-10-28	15
Individual 8	A125-006	7	F	2017-10-13 to 2017-10-28	13
Individual 9	A134-005	8	M	2017-10-16 to 2017-10-26	10
Individual 10	A134-006	6	F	2017-10-13 to 2017-10-30	16
Individual 11	A140-001	25	F	2017-10-13 to 2017-10-25	13
Individual 12	A177-002	51	F	2017-10-13 to 2017-10-26	9
Individual 13	A177-003	11	F	2017-10-13 to 2017-10-23	6
Individual 14	A177-004	2	F	2017-10-13 to 2017-10-24	7
Individual 15	K104-002	43	F	2017-10-10 to 2017-10-28	14
Individual 16	K113-002	13	M	2017-10-19 to 2017-10-30	12
Individual 17	K113-005	40	M	2017-10-19 to 2017-10-28	15
Individual 18	K119-008	12	F	2017-10-11 to 2017-10-30	12
Individual 19	K119-009	10	M	2017-10-11 to 2017-10-30	18
Individual 20	K138-004	16	M	2017-10-10 to 2017-10-30	12

Table 6.1 shows the summary of datasets used in this analysis. The number of days available for analysis are selected as any day in the experiment where contact events are logged. All the days in the experiment that had no contact events have been removed. The first three digits of the Study Id, following the letter A or K, indicates the household number. As seen from Table 6.1, there is not complete household data, except for household A112. As a result, the analysis is done at an individual level.

6.2 Data Preparation and Cleaning

The individual datasets are pre-processed before principal component analysis is carried out. This includes normalizing the data such that each day carries a weight of 1, otherwise some days will dominate the principal components and misrepresent the contact events. Secondly, the days that have a single contact event, have also been removed, because that single contact event will carry a weight of 1 and also

dominate the principal components, which will bias the results. This is also justified, because a day that has a single contact event, can be seen as an outlier, thus removed from the analysis.

6.2.1 Data Cleaning Results

A false contact event occurs when two sensors are in close proximity, but are not worn by individuals. A true contact event is assumed to be a logged contact event and with a corresponding status value indicating movement of the sensors. This movement of the sensors is assumed to represent the sensor being worn by the individual and thus, any contact event logged in this scenario is assumed to be true. A status flag represents the state of the sensor i.e. worn or not by flag values 16 and 0, respectively. A contact event and status matrix is constructed, for the number of days of available data. A Haddamard product which is an element-wise multiplication of the matrices, produces a resulting matrix of true contact events. This resulting matrix is used as the input matrix for the principal component analysis. Figure 6.1 shows the raw data for Individual 1. Similar plots for all other individuals are attached in Appendix D.

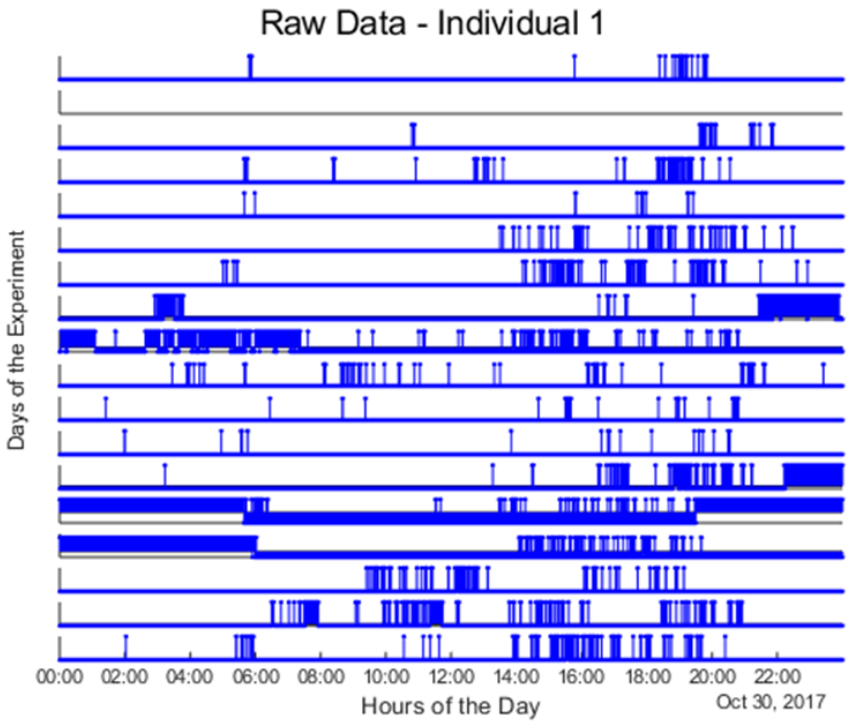


Figure 6.1: Raw Data of Individual-1. Each subplot represents a day in the experiment. Y-axis shows the contact events and x-axis shows the time of day.

Figure 6.1, shows the raw data for Individual 1. This represents all the logged contact events by the sensor allocated to Individual 1. The stems represent a contact event between the sensor belonging to Individual 1 and another sensor. The dense areas represent the false contact events or noise that represent two sensors in close proximity, but are not worn by the individuals.

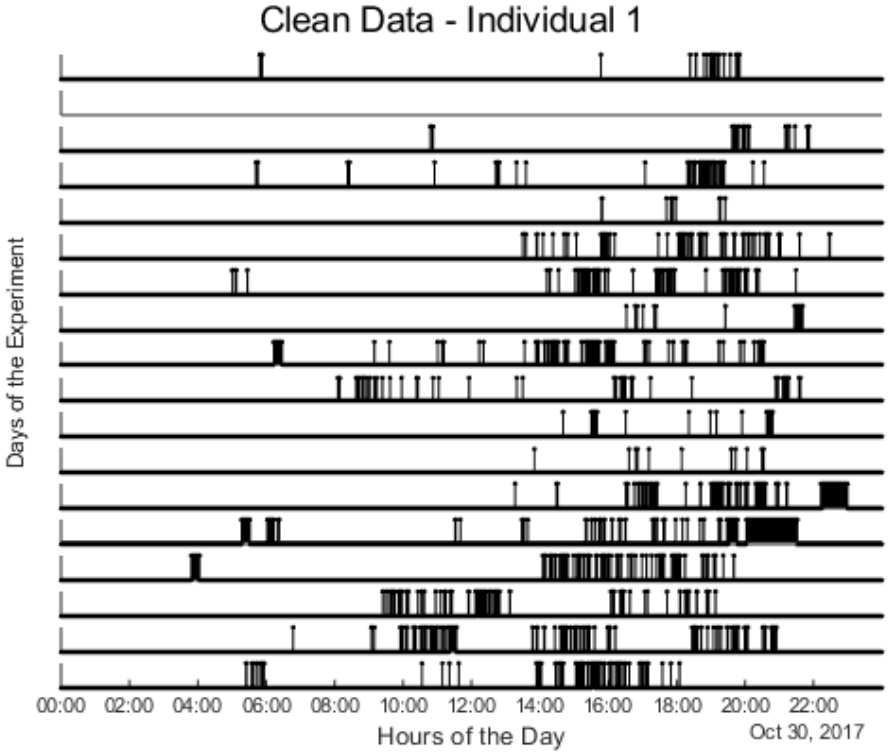


Figure 6.2: Cleaned Data for Individual 1. Each subplot represents a day in the experiment. Y-axis shows the contact events and x-axis shows the time of day.

Figure 6.2, shows the cleaned data for Individual 1. It can be seen that the dense areas have been removed, by using an element-wise multiplication of the contact matrix and status matrix. All the resulting contact events shown in Figure 6.2 are those that have been logged by the sensors in motion, which indicates that it has been worn by the individual, at the time of contact event logging. These are the true contact events used for analysis.

6.2.2 Percentage Usable Data

Percentage usable data is a measure of the proportion of data that represents true contacts i.e. it is a percentage of cleaned data from the overall contact events. This is computed using the following formula.

$$Percentage\ Usable\ Data = \frac{Cleaned\ Data}{Original\ Data}\ x\ 100$$

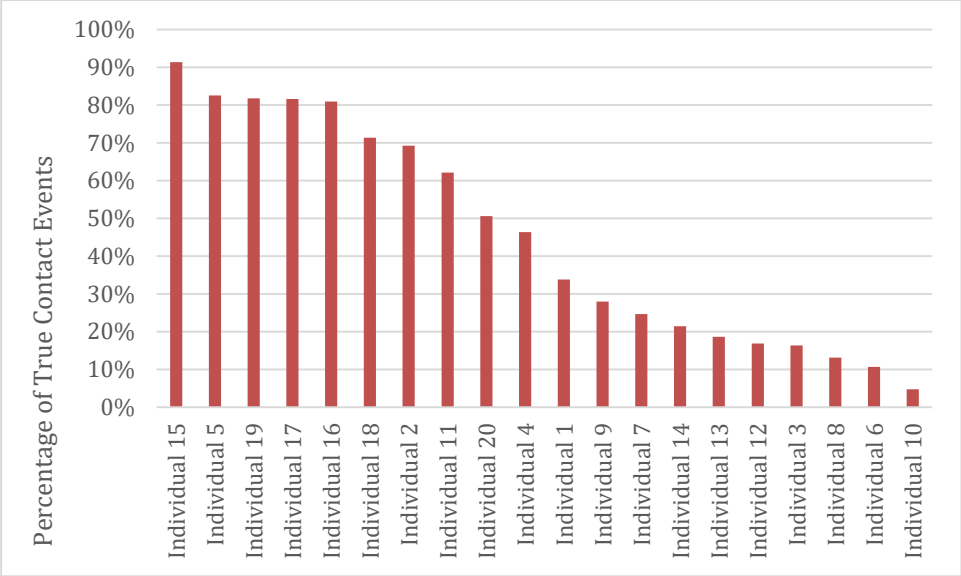


Figure 6.3: Percentage Usable Data of Logged Contact Events. The y-axis shows the percentage proportion of cleaned data over the original data. The x-axis shows all the individuals

The proportion of true contact events gives a measure of how much data of this kind is useable. In other words, when proximity sensors are deployed in a study in South Africa, how much of true contact events do you get? As seen from Figure 6.3, this varies across the individuals.

6.3 Selection of Sampling Period

The sampling period is important to represent the contact events over a specified temporal scale. Contact events are logged every 20-seconds, over a 24-hour period. Five sampling periods are considered to determine which sampling period provides the best representation of the variance in the data. These sampling periods are: 20-

seconds, 5-minutes, -10-minutes, 20-minutes, 1-hour. This is done by sampling the data at the various sampling periods and computing the cumulative variance representation of the first three principal components. The first three principal components are chosen, because they can be easily visualised. Figure 6.4, shows a summary of the cumulative variance representation per sampling period chosen

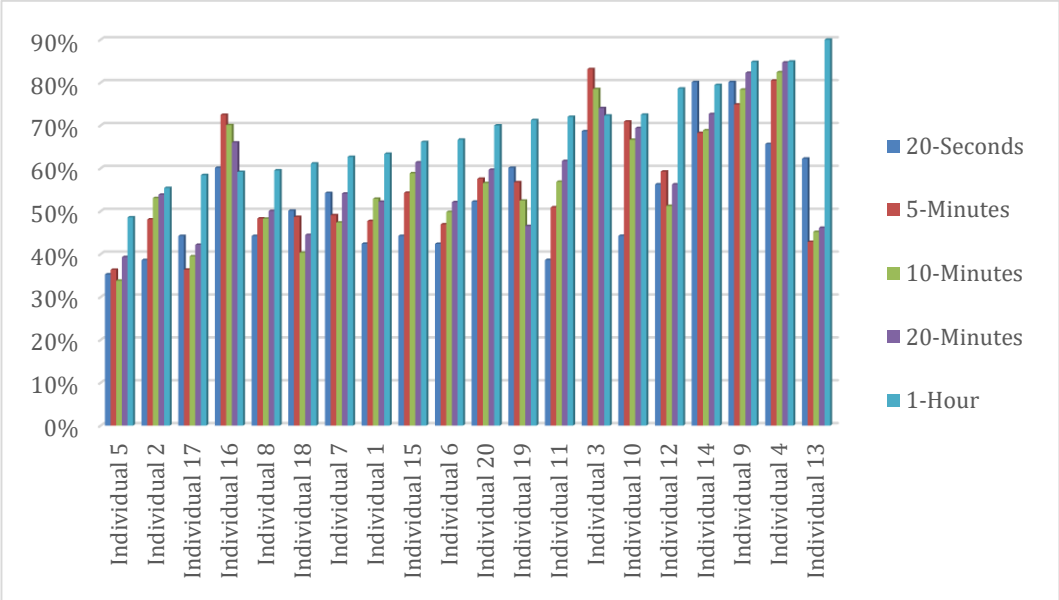


Figure 6.4: Cumulative Variance Representation by the First Three Principal Components for Each Sampling Period. The y-axis shows the percentage cumulative variance of the first three principal components, sorted in ascending hour by various representation of 1-hour sampling period. The x-axis shows the individuals in the study. The cumulative variance is for each sampling period is represented by the various bars on the plot.

Figure 6.4, shows the cumulative variance representation of the first three principal components. This conveys how much variance of the original data is retained when considering only the first three principal components, using the five sampling periods considered. From Figure 6.4, the amount of variance representation increases with an increase in sampling period i.e. a smaller sampling period like 20-seconds, gives the least representation of variance, as compared to a higher sampling period of 1-hour. There are exceptions with individual 3 and 16, where the sampling period of 5-minutes yields a similar and higher variance representation to the 1-hour sampling period, respectively. In order to make the datasets

comparable, one sampling period is chosen that gives the best representation of variance and that is 1-hour. There will be 24 samples for each time trends (i.e. day of the experiment).

6.4 Presentation of Results

The twenty individuals have been grouped into 4 Classes based on the measurement of contact events. Class-1 refers to a group of individuals with consecutive days of contact activity, without a loss of data or gap between the days. The remaining Classes do not have consecutive days of measurement. Class-2 refers to a group of individuals with contact events for seven days or less. Class-3 refer to individuals with contact events for at least 8 days and less than 14 days. Class-4 refers to individuals with contact events of at least 14 days and more. A comparison of all dataset in a class is then summarised in a table at the end of each section. Appendix E is used as a reference throughout this section.

6.4.1 Mapping the Variables to the Calendar Days

There are twenty individual datasets with contact events logged for a period ranging from six to 18 days. Table 6.2 shows a calendar grid that is used to map the days that are the variables in this analysis, to the actual calendar day of the experiment. In Table 6.2, the cells that are highlighted in pink show the days that have been removed due to no contact events logged for that day. The cells highlighted in green represent the days that have been removed, as a result of only a single contact event logged that day and those days will carry the most weight, ultimately dominating the principal component analysis, which will bias the results and misrepresent the contact patterns.

Table 6.2: Calendar Grid for 1-to-1 Mapping Between Day Number and Calendar Day

Date	Calendar Day	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7	Ind 8	Ind 9	Ind 10	Ind 11	Ind 12	Ind 13	Ind 14	Ind 15	Ind 16	Ind 17	Ind 18	Ind 19	Ind 20
10-10-2017	Tuesday															D1		D1	D1	D1	D1
11-10-2017	Wednesday																		D2	D2	D2
12-10-2017	Thursday															D2			D3	D3	D3
13-10-2017	Friday	D1	D1	D1		D1	D1	D1			D1	D1	D1	D1	D1	D3			D4		D4
14-10-2017	Saturday					D2	D2					D2				D4		D2	D5	D4	
15-10-2017	Sunday	D2	D2				D3				D2	D3	D2	D2	D2				D6	D5	
16-10-2017	Monday	D3	D3	D2		D3	D4	D2		D1	D3	D4	D3	D3	D3			D3	D7	D6	
17-10-2017	Tuesday	D4				D4	D5	D3	D1	D2	D4	D5		D4				D4	D8		D5
18-10-2017	Wednesday	D5			D1	D5	D6	D4	D2	D3	D5	D6	D4	D5	D4				D9		D6
19-10-2017	Thursday	D6	D4	D3		D6	D7	D5	D3	D4		D7	D5	D6		D5	D1	D5	D10	D7	
20-10-2017	Friday		D5				D8	D6	D4		D6	D8	D6		D5	D6	D2	D6	D11	D8	D7
21-10-2017	Saturday	D7	D6	D4	D2		D9	D7		D5	D7	D9	D7		D6	D7	D3	D7	D12	D9	
22-10-2017	Sunday	D8	D7	D5		D7	D10	D8	D5	D6	D8	D10				D8	D4	D8		D10	D8
23-10-2017	Monday	D9	D8			D8	D11	D9	D6	D7	D9	D11				D9	D5	D9		D11	
24-10-2017	Tuesday	D10	D9			D9		D10	D7	D8	D10	D12			D7	D10	D6	D10		D12	D9
25-10-2017	Wednesday	D11	D10	D6	D3	D10	D12	D11	D8	D9	D11		D8			D11	D7	D11		D13	
26-10-2017	Thursday	D12	D11		D4	D11	D13	D12	D9	D10	D12					D12	D8			D14	
27-10-2017	Friday	D13	D12		D5	D12		D13	D10		D13					D13	D9	D12		D15	D10
28-10-2017	Saturday	D14		D7	D6	D13	D14	D14	D11		D14					D14	D10	D13		D16	D11
29-10-2017	Sunday	D15		D8	D7	D14	D15	D15	D12		D15						D11	D14		D17	
30-10-2017	Monday	D16	D13	D9		D15			D13		D16						D12	D15		D18	D12

From Table 6.2, Individual 11, 16 and 18 have consecutive days of contact activity, 12 days in total. The compliance of the participants wearing the sensors can also be observed. The more the number of days of no contact activity are observed, the less compliant the individuals are to the study. Some behavioural patterns can also be observed from this calendar grid e.g. most of the individuals did not wear the sensors on Saturday, 14-10-2017 or that the individuals are generally not in the household on most weekends. However, the importance of this grid is to compare the similarities between the calendar days of the month.

6.5 Class-1: Consecutive Days of Contact Activity

This category of refers to individual 11, 16 and 18, who all have consecutive days of contact activity. All three individuals have 12 days of logged contact activity. The full analysis for individual 11 is presented and then a summary table is included for all individuals in this category. This will determine how the consecutive days of contact events vary over time.

6.5.1 Visual Inspection of Contact Events – Time Trends

In this section, the visual inspection of the time trends is presented. Each variable i.e. day of the experiment, is a measure of contact events over 24-hours.

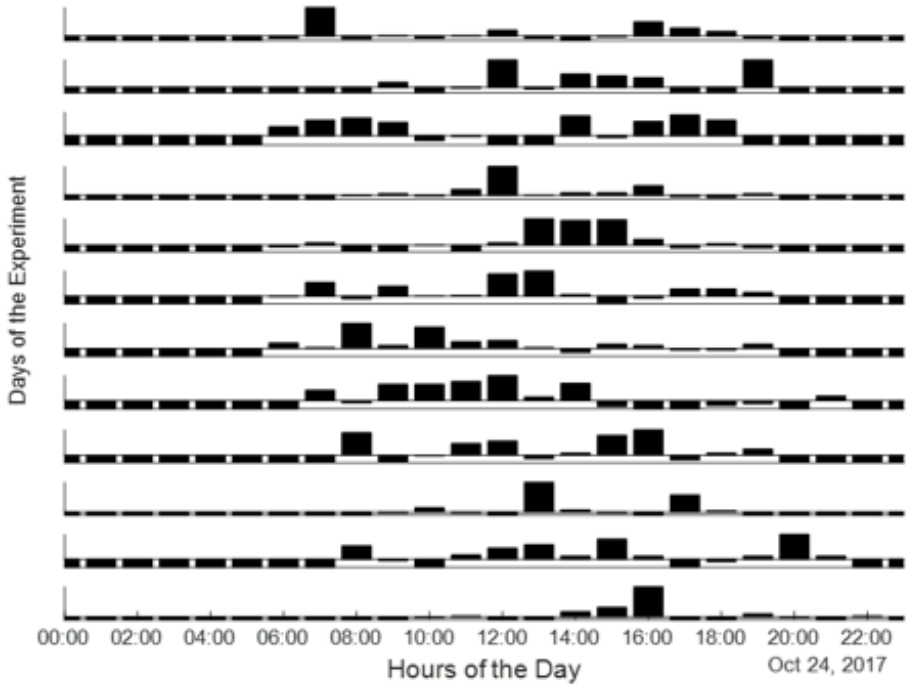


Figure 6.5: Contact Events over Time for Individual 11. The subplots i.e. y-axis shows the days of the experiment (top subplot is day one and bottom subplot is the last day) and the x-axis shows the hour of the day. Each bar represents contact events for that hour.

Contact events for Individual 11 occur in the time range between 07:00 and 21:00. Day 2, 4 and 5 are similar, because contact events are consistently logged at similar times, even though the magnitude or intensity of contact events may differ slightly. Day 2 and Day 10 also show similarities in terms of timing of contact events. Day 6, 7 and 8 are also similar in terms of the time of logged contact events. Day 9 and 11 are also similar. Day 1, 3 and 12 differ from the rest of the days, in terms of the time of logged contact events. From Individual 11, three clusters of similar days are observed.

6.5.2 Scree Plot

The scree plots of all individuals in Class-1 are found in Appendix E.

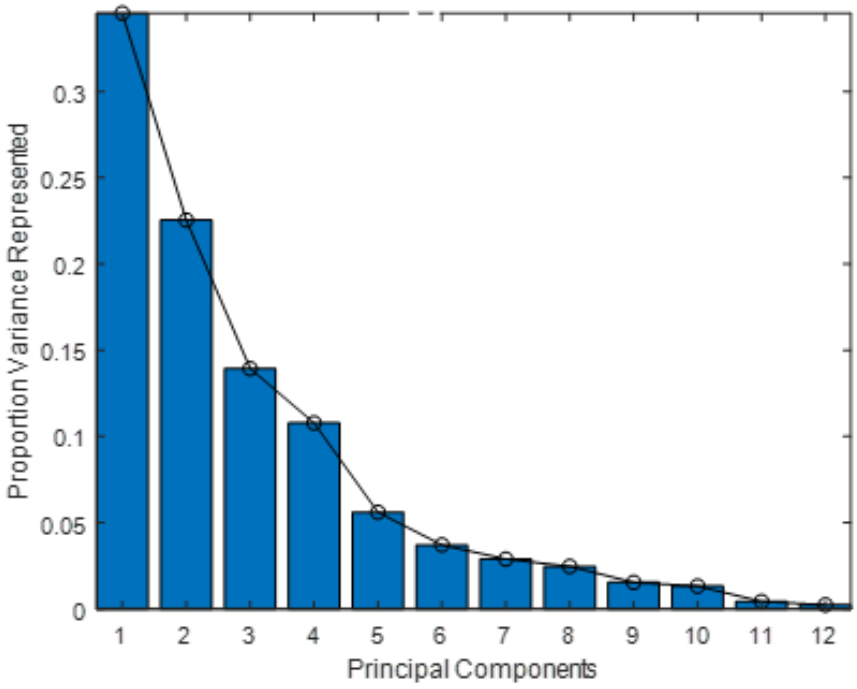


Figure 6.6: Scree Plot Showing Proportion Variance Represented by Each Principal Component

An example is provided in Figure 6.6, this plot shows that the first principal component for Individual 11 accounts for 35% of the variation. The addition of the second principal component adds 23% of the variation. The cumulative variance representation by the first two and three principal components is 58% and 72% respectively. Choosing the first three principal components for Individual 11 will adequately represent the original variance, without significant loss of information in the original data. The first three principal components will be used to represent the contact events for Class-1.

6.5.3 Principal Component Loadings

Table 6.3: PC Loadings for Individual 11

PC Loadings - Individual 11				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,1237	0,2404	0,0874	**0,081**
Day 2	0,3078	0,2134	0,1998	**0,180**
Day 3	0,0902	0,1306	-0,0074	**0,025**
Day 4	0,2421	0,3501	0,2161	**0,228**
Day 5	0,3060	0,1283	-0,2739	**0,185**
Day 6	0,0834	0,3275	-0,1649	**0,141**
Day 7	0,1085	0,1461	0,0269	**0,034**
Day 8	0,0599	0,3037	0,0286	**0,097**
Day 9	0,2958	0,1015	0,1702	**0,127**
Day 10	0,0536	0,3837	-0,7310	**0,684**
Day 11	0,1984	0,0441	-0,0539	**0,044**
Day 12	0,4022	0,0786	0,3885	**0,319**

Table 6.3, shows the contribution of each variable to the first three principal components. Each loading is considered a coordinate in the loadings plot. The magnitude or contribution of each variable is calculated as the square sum of the loadings for each principal component, given in the last column.

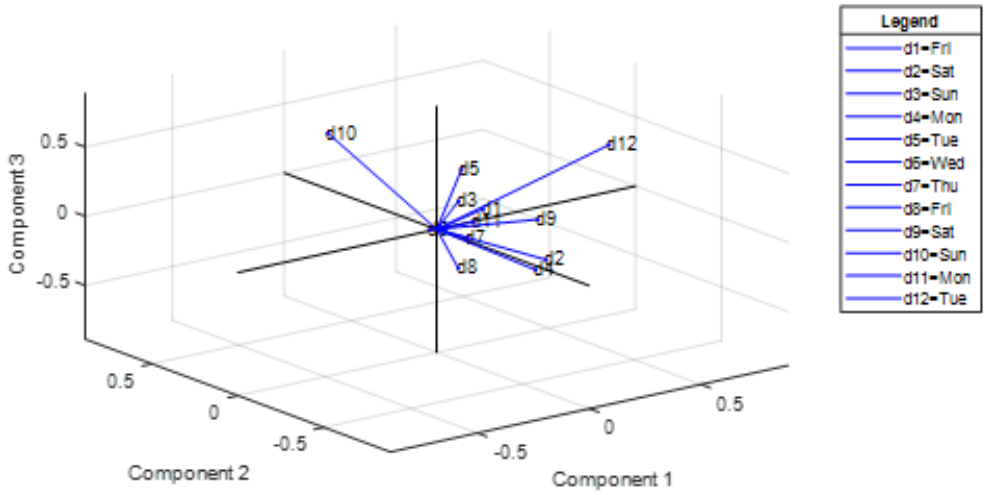


Figure 6.7: Loadings Plot for Individual 11. This shows the span of the original variables in the first three principal components space for Individual 11

From Figure 6.7, day 8, day 10 and day 12, are different from all the days. Two clusters of days can be observed, cluster one = day 2, day 4, day 7 and day 9; cluster two = day 1, day 3, day 5 and day 11.

6.5.4 Comparison of Clusters

From the visual inspection of time trends, there are three clusters observed of similar days, compared to two clusters of similar days from the loadings plot. The only consistency between the two visualisations are day 2 and day 4. The difference between these methods is that the visual inspection looks at the similarity in the timing of contact events i.e. how repetitive are the contact events over the days. Whereas, the loadings plot looks at the variation of the contact events in terms of the intensity or the weight of contact events, across the days, to determine the similarity between the days.

6.5.5 Significance of Days to Principal Components

Each principal component can be represented as a linear combination of the original variable, by expanding (17) from Section 5 i.e.

$$PC = \sum_{i=1}^{N} \alpha_i \, x_i$$

Where α_i terms are given by the contribution column in Table 6.3 and x_i terms represent the days of contact activity. The representation of the principal components as a linear combination of the original variables makes it easier to quantitatively determine, which variables are important towards the first three principal components and which are redundant. This is an important part of feature selection. Depending on a chosen threshold value, variables that meet the criteria can be regarded as significant. For example, if variables with a contribution of at least 10% are retained, then from 17 and Table 6.3:

For Individual 11: day2, day4, day 5, day 6, day 9, day 10, day 12 will be important.

6.5.6 Comparison of Clusters and Significant Days to Calendar Days

In this section, the clusters and significant days that have a strong contribution to the principal components are compared to the original variables that represent the calendar days.

According to the loadings plot for Individual 11, day 2, 4, 7 and day 9 form a cluster. These days represent Saturday, Monday, Thursday and Saturday, respectively. The second cluster is formed by day 1, day 3, day 5 and day 11. These days represent Friday, Sunday, Tuesday and Monday. Even though contact events for 12 consecutive days of measurement are available for Individual 11, it shows that there is still clustering between these days, which means that some days are similar and some are different. The results also show that the clustering is observed across calendar days that are far apart. For a given threshold of 10%, the significant days for individual 11 are day 2 = Saturday; day 4 = Monday; day 5 = Tuesday; day 6 = Wednesday; day 9 = Saturday; day 10 = Sunday and day 12 = Tuesday. This shows that these days do carry important information to represent Individual 11's contact activity. The more significant the days are in the principal component space; the more variation they represent. This is also a measure of how important those days are in representing the contact activity of the individual. Therefore, Individual 11

has at least seven days that have a significant contribution to the first three principal components and a measurement of at least seven days is required to represent Individual 11's contact activity.

6.5.7 Summative Comparison of Class-1 Individuals

Table 6.4: Summary of Main Results for Class-1

Individual ID	Cumulative Variance	Clusters of Similar Days		Number of Clusters: Loadings Plot	Number of Outliers: Loadings Plot	Number of Significant Days
		Time Trends	Loadings Plot			
Individual 11	72%	{d2,d4,d5,d10}; {d6,d7,d8}; {d9,d11}	{d2,d4,d7,d9}; {d1,d3,d5,d11};	2	3	7
Individual 16	59%	{d1,d12}; {d5,d7,d8}; {d9,d10,d11}	{d1,d2,d5,d8,d12}; {d3,d4,d9,d11}	2	3	9
Individual 18	61%	{d1,d2,d8}; {d3,d4,d5,d6}; {d9,d10}	{d1,d8}; {d7,d10}; {d5,d11}; {d2,d4,d6,d12}; {d3,d9}	5	0	9

Table 6.4, shows a summary of the analysis done on Class 1. The table shows that there is a clustering of days that are similar for each Individual. The table also shows depending on the method used, different clusters may be observed. This table conveys that the contact events vary over the days and some days may be similar, but not all the days are the same. The higher the number of clusters, the greater the variation in contact activity across the days. This table shows that for these individuals, at least 7 days for Individual 11, and 9 days for Individual 16 and Individual 18, are required to adequately represent the contact activity, based on a minimum threshold of 10%.

From this point onwards, the main findings from the remaining Class will be presented and the plots will be referenced from Appendix E.

6.6 Class-2: Contact Events for Seven Days or Less

In this section, the results of individuals who have contact events logged for seven days or less are presented. In this category, the analysis of Individual 4, Individual 13 and Individual 14 are presented. The focus will be on Individual 4, then a summative comparison will be given in a Table for all Individuals in this category.

6.6.1 Visual Inspection of Contact Events – Time Trends

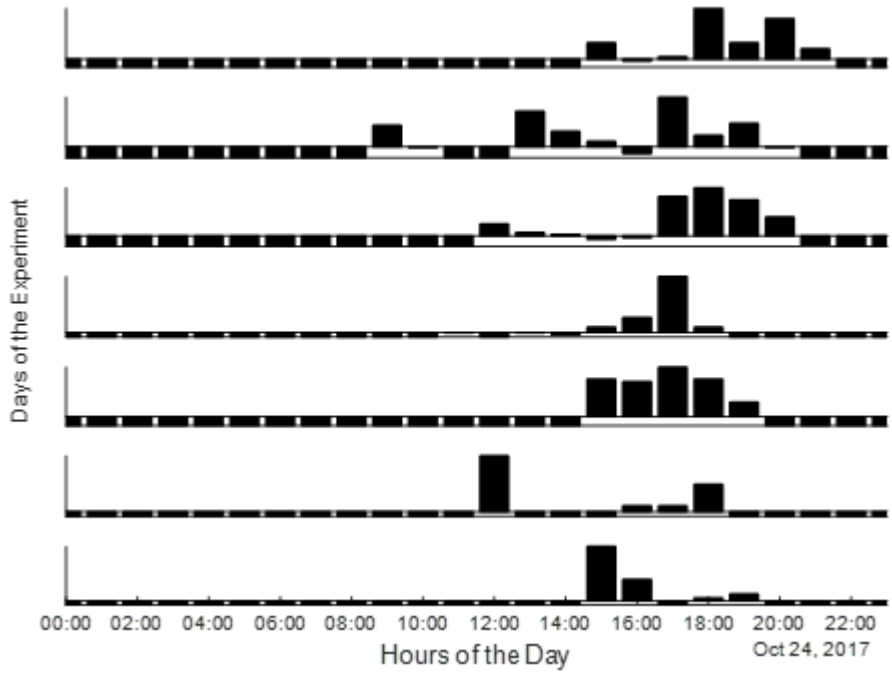


Figure 6.8: Contact Events for Individual 4. The subplots i.e. y-axis shows the days of the experiment (top subplot is day one and bottom subplot is the last day) and the x-axis shows the hour of the day. Each bar represents contact events for that hour.

Following Figure 6.8, it can be observed that the contact events for Individual 4 are fairly consistent over the days. Days 1, 2 and 3 are similar. Day 4 and 5 are similar. Day 6 and 7 are different from the other days, but could also form clusters. There are two cluster that can be observed for Individual 4.

6.6.2 Scree Plot

The scree plots for Class 2 are found in Appendix E. The cumulative variance representation by the first three principal components is 85%, 90% and 79% for

Individual 4, Individual 13 and Individual 14, respectively. This is a very good representation of variance.

6.6.3 Principal Component Loadings

The tables for the principal component loadings are found in Appendix E. In this section, the loadings plot and clustering of days is presented for Individual 4.

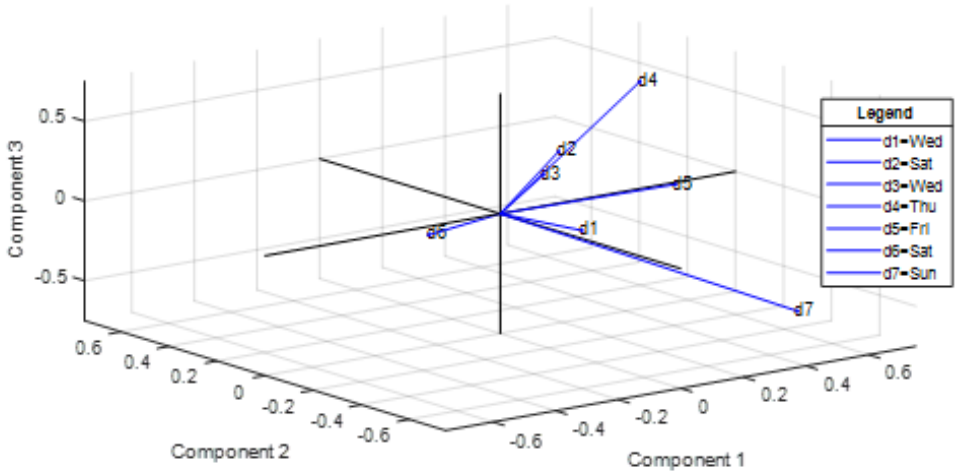


Figure 6.9: Loadings Plot for Individual 4. This shows the span of the original variables in the first three principal components space for Individual 4.

Following Figure 6.9, there are two main clusters and two outliers. Day 2, 3 and 4 are similar. Day 1 and day 7 are similar, even though they don't have the same contribution. Day 5 and day 6 seem to be outliers. However, Day 4 and Day 5 span in the same direction of principal component 1, but day 4 also has a contribution to principal component 3, where day 5 does not. These days could also be seen as similar.

6.6.4 Comparison of Clusters

From the visual inspection of the time trends and the loadings plots, day 2 and 3 are consistent in similarity. Day 1 and 7 are not entirely similar from the time trends, but they do have contact activity around the same time. This also reflects in the loadings plot, both days span in the same direction, but are not on the same point in space, which makes them slightly different in terms of their contributions. Day 4 and Day 5 are similar in the time trends and span in the same direction of principal

component 1, but not principal component three. The differences in grouping by the two methods could be that the intensity of contact events is difficult to visualise using only the time trends. The loadings plots give a good quantitative representation of the similarity and differences between the days.

6.6.5 Significance of Days to Principal Components

The principal component loadings tables for Class 2 are found in Appendix E. Using principal component loadings table and a threshold of 10%, then all the days for Individual 4 will be important.

6.6.6 Comparison of Clusters and Significant Days to Calendar Days

Two clusters from the loadings plot are observed for Individual 4. Cluster one is formed by day 2, day 3 and day 4. These days represent Saturday, Wednesday and Thursday. Even though these days seem far apart, except for Wednesday and Thursday, the contact activity across these days is similar. The second cluster is formed by day 1 and day 7, representing Wednesday and Sunday respectively. These days are also far apart, but their similarity can be grouped together. With a threshold of 10%, all the days for Individual 4 are significant in the principal component space. This means that none of the days are redundant and that at least seven days are required to adequately represent Individual 4's contact activity.

6.6.7 Summative Comparison of Class-2 Individuals

Table 6.5: Summary of Main Results for Class-2

Individual ID	Cumulative Variance	Clusters of Similar Days		Number of Clusters:	Number of Outliers:	Number of Significant
		Time Trends	Loadings Plot			
Individual 4	85%	{d1,d2,d3}; {d4,d5}; {d6,d7}	{d2,d3,d4}; {d1,d7}	2	2	7
Individual 13	90%	{d1;d3}	{d1,d2,d3,d6}; {d4,d5}	2	0	5
Individual 14	79%	{d1,d3,d5}; {d4,d6}	{d1,d2,d3,d5,d6}	1	2	5

Table 6.5, shows a summary of the results for Class 2. These groups of individuals have contact events logged for 7 days or less. Class 2 shows a consecutive similarity

in contact activity. For example, Individual 4 has similar days between day 1-3 and day 4-6 and day 6-7, even though these are not consecutive calendar days. This group show a smaller number of clusters across these days and have more days forming a cluster as compared to Class 1. All the days for Individual 4 are significant in the principal component space. All Individuals shows a great deal of clustering across the days, which shows that not all the days are the same, but some days do exhibit similar contact activity. Class 2 also have the highest representation of variance, which means that the smaller dimensions of three principal components, really represents the original variance in the data very well.

6.7 Class-3: Contact Events for 8-13 Days

In this section, the results for individuals with logged contact events for 8-13 days are presented. In this category, the results for Individual 2, Individual 3, Individual 8, Individual 9, Individual 12 and Individual 20, are presented. The focus of this section will be on the contact activity of Individual 2, then a summative comparison across all other individuals in this category will be presented in a table.

6.7.1 Visual Inspection of Contact Events

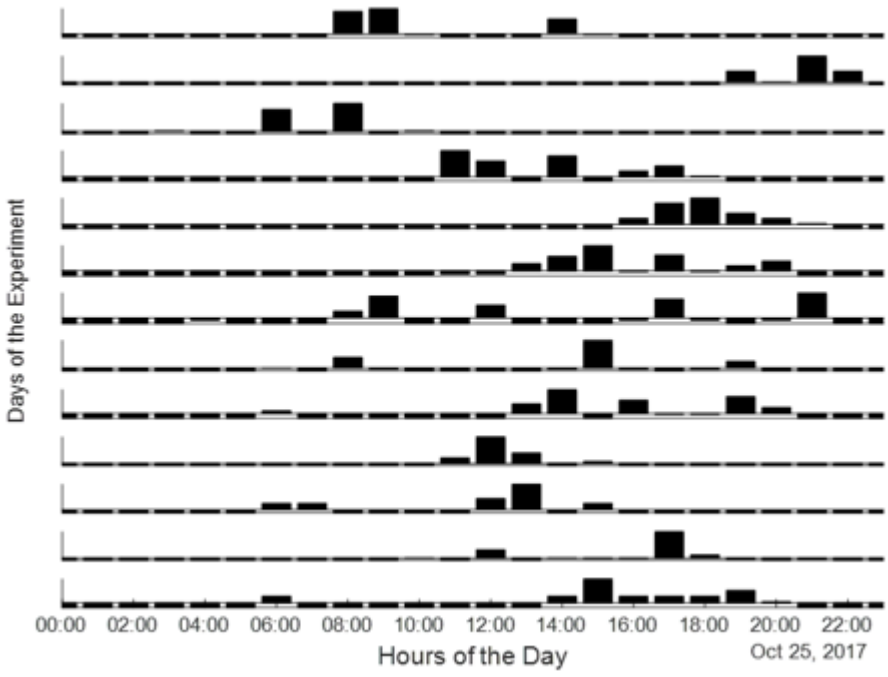


Figure 6.10: Contact Events over Time for Individual 2. The subplots i.e. y-axis shows the days of the experiment (top subplot is day one and bottom subplot is the last day) and the x-axis shows the hour of the day. Each bar represents contact events for that hour.

Following Figure 6.10, there are four clusters of similar days: cluster one = day 1 and 3; cluster two = day 5 and day 6; cluster three = day 7, day 8, day 9 and day 13; cluster four = day 10 and 11. Day 2 and Day 12 seem different from rest of the days.

6.7.2 Scree Plot

The scree plots for Class 3 are found in Appendix E. The cumulative variance representation of the first three principal components is: Individual 2 (55%), Individual 3 (72%), Individual 8 (60%), Individual 9 (85%), Individual 12 (79%) and Individual 20 (70%). The first three principal components will represent the data without much loss of information in the original data.

6.7.3 Principal Component Loadings

The tables for the principal component loadings are found in Appendix E. In this section, the loadings plot and clustering of days is presented for Individual 4.

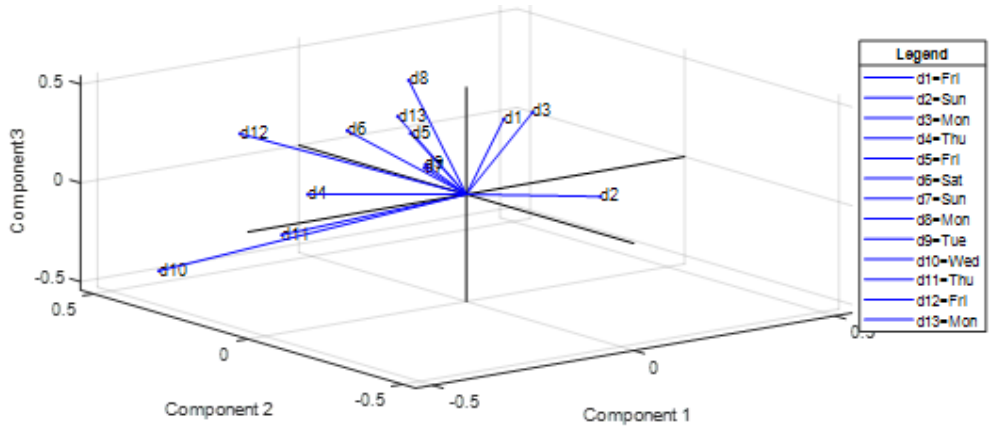


Figure 6.11: Loadings Plot for Individual 2. This shows the span of the original variables in the first three principal components space for Individual 2

Following Figure 6.10, shows the loadings plot for Individual 2. This plot shows that many variables form a cluster. There are three main clusters: cluster one = day 1 and day 3; cluster two = day 4, day 11 and day 12; cluster three = day 5, day 6, day 7, day 8, day 9 and day 13

6.7.4 Comparison of Clusters

From the visual inspection of the time trends, four clusters were observed, as compared to the three clusters from the loadings plot. Day 1 and day 3 are consistent in both plots. Day 5 and 6; day 7, 8, 9, and 13 are also consistent in both plots. Day 5 and 6 are observed as one cluster from the time trends, but are a part of a bigger cluster of similar days from the loadings plot. Day 2 and day 12 are outliers in both plots.

6.7.5 Significance of Days to Principal Components

The principal component loadings tables for Class 3 are found in Appendix E.

With a threshold of 10%, for Individual 2, day 1, day 2, day 3, day 4, day 5, day 6, day 8, day 10, day 11, day 12 and day 13 are significant in the first three principal components i.e. a total of 11 days are important.

6.7.6 Comparison of Clusters and Significant Days to Calendar Days

Three clusters are observed from the loadings plot for Individual 2. Cluster one is day 1 and day 3, these represent Friday and Monday, respectively. The second cluster is formed by day 4, day 10 and day 11; these represent Thursday, Wednesday and Thursday. Even though these days are not consecutive, but represent close calendar day, it shows that there is a repeatability of contact activity for some of the calendar days. The third cluster is formed by day 5, day 6, day 7, day 8, day 9 and day 13, which represent Friday, Saturday, Sunday, Monday, Tuesday and Monday. These are the last days of the experiment and fortunately, the calendar days were consecutive. This shows that for Individual 2, there is great deal of similarity across the days and that some of the calendar days have similar

contact activity across the experiment period. A total of 11 days are significant in the first three principal component space.

6.7.7 Summative Comparison of Class-3 Individuals

Table 6.6: Summary of Main Results for Class 3

Individual ID	Cumulative Variance	Clusters of Similar Days		Number of Clusters: Loadings	Number of Outliers: Loadings	Number of Significant Days
		Time Trends	Loadings Plot			
Individual 2	55%	{d1,d3}; {d5,d6}; {d7,d8,d9,d13}; {d10,d11}	{d1,d3}; {d4,d10,d11}; {d5,d6,d7,d8,d9,d13}	3	2	11
Individual 3	72%	{d1,d2,d6}; {d3,d4}	{d3,d4}; {d1,d2,d6}	2	4	5
Individual 8	60%	{d1,d2}; {d3,d4,d5}; {d7,d8}; {d9,d10,d11}	{d1,d2,d3,d4,d7,d8,d9,d10}; {d5,d11}; {d12;d13}	3	1	4
Individual 9	85%	{d1,d2,d3,d7,d10}; {d4,d5,d6}; {d8,d9}	{d1,d2,d3,d7}; {d8,d9}; {d4,d5,d6,d10}	3	0	10
Individual 12	79%	{d1,d2,d7,d8}; {d3,d6}; {d4,d5}	{d3,d6}; {d1,d2,d7}	2	3	8
Individual 20	70%	{d1,d2,d3,d5,d6,d7,d8,d10}; {d8,d2}	{d3,d6}; d2,d5,d8}; {d1,d3,d7,d9}; {d11,d12}	4	1	10

Table 6.6, shows the summary of the results for Class 3. This group of individuals have contact events that range from 8 – 13 days. This group of individuals also demonstrate a clustering of contact events. This group of individuals have calendar days that have similar contact events shown in the clustering of the days. Individual 2 and Individual 8 have a large cluster of similar days. Individual 9 has three clusters of similar days without a single outlier. From this group, it seems that the more the number of logged contact days are available, the greater the clustering of similar days. All the individuals in this group have at least 8 days of significant days in the principal component space, except for Individual 3 and five. The smaller the number of significant days in the principal component space, in proportion to the number of original days of measurement, the greater the similarity of contact activity, which is shown by the clustering of the days.

6.8 Class-4: Contact Events or 14 Days or More

In this section, the results for Individual with logged contact events for 14 days or more. In this category, the results for Individual 1, Individual 5, Individual 6, Individual 7, Individual 10, Individual 15, Individual 17 and Individual 19, are presented. The focus of the analysis will be on Individual 2, then a summary of the results for all the individual will presented in a table. All the plots for this section are referred to Appendix E.

6.8.1 Visual Inspection of Contact Events – Time Trends

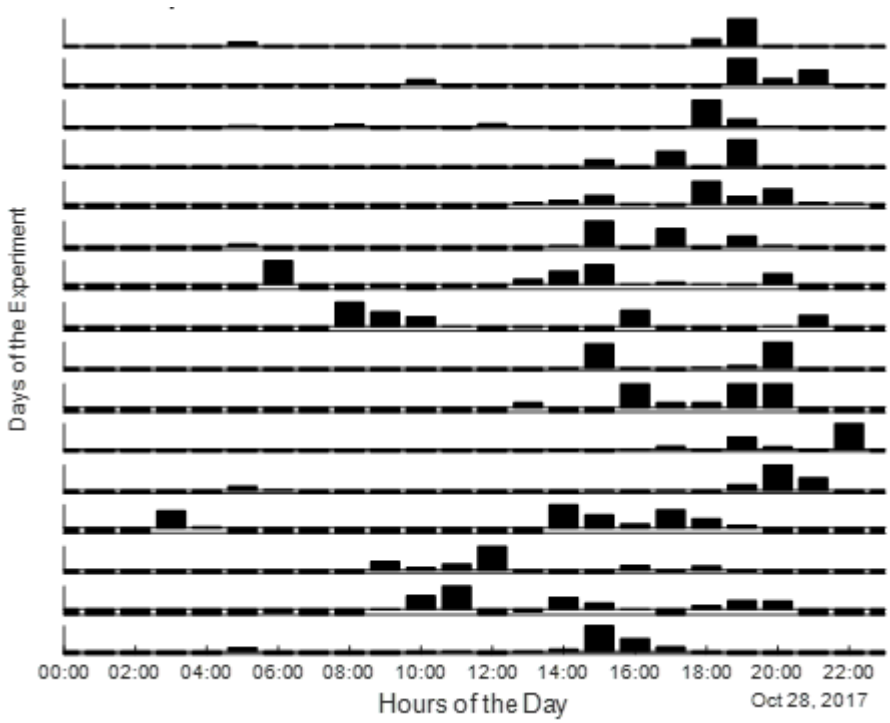


Figure 6.12: Contact Events over Time for Individual 1. The subplots i.e. y-axis shows the days of the experiment (top subplot is day one and bottom subplot is the last day) and the x-axis shows the hour of the day. Each bar represents contact events for that hour.

Figure 6.12, shows the contact events over time for Individual 1. There are 16 days of logged contact events. There are four main clusters of similar days observed for Individual 1. In fact, there is only one outlier i.e. day 16, and the rest of the days form clusters of similar contact activity. A lot of the contact activity takes place

consistently in the evenings for Individual 1. Cluster 1 is formed by day 1, day 2, day 3, day 11 and day 12. Cluster 2 is formed by day 4, day 5, day 6, day 9 and day 10. Cluster 3 is formed by day 7, day 8 and day 13. Cluster 4 is formed by day 14 and day 15.

6.8.2 Scree Plot

The scree plot for Individual 1 shows that the cumulative variance representation of the first three principal components is 63%, which is good enough to represent the original variance in the data without significant loss of information. The scree plots are attached in Appendix E.

6.8.3 Principal Component Loadings

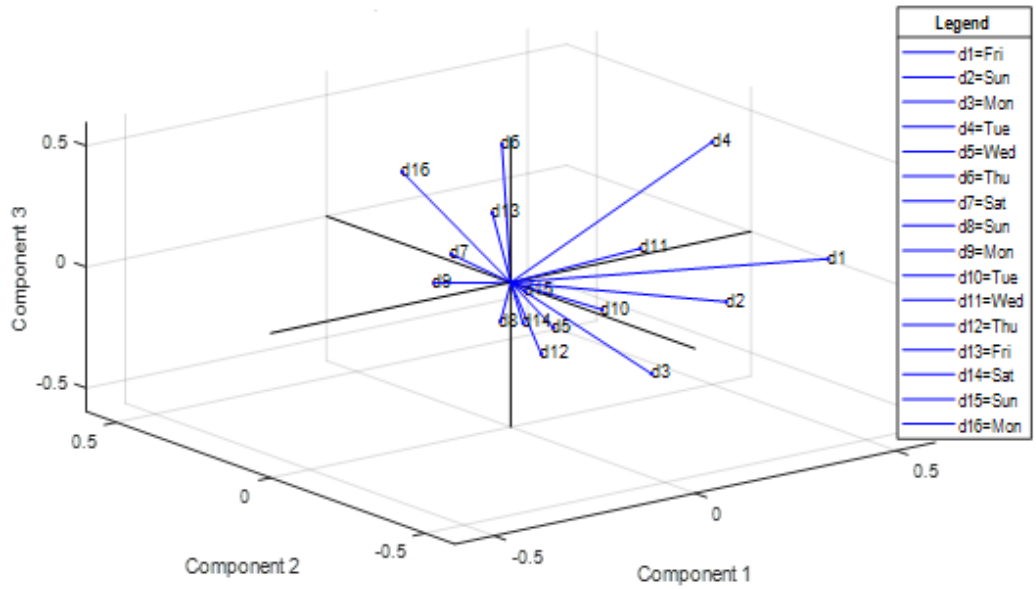


Figure 6.13: Loadings Plot for Individual 1. This shows the span of the original variables in the first three principal components space for Individual 1

From the loadings plot in Figure 6.13, four clusters of similar days are observed. Cluster 1 is formed by day 1, day 2 and day 11. Cluster 2 is formed by day 3, day 5, ay 8, day 10, day 12, day 14 and day 1. Cluster 3 is formed by day 7, day 9 and day 13. Cluster 4 is formed by day 6 and day 16.

6.8.4 Comparison of Clusters

There are consistencies in the number of clusters observed from the time trends and the loadings plot. Day 1, day 2 and day 11 are similar in both plots. Day 5 and day 10, as well day 7 and 13 are presented in similar clusters from both plots. The contact activity shows similarity across the days, with some days being more similar than others

6.8.5 Significance of Days to Principal Components

The principal component loadings for Class 4 are found in Appendix E. With a threshold of 10%, the following days contribute significantly to the first three principal components, for Individual 1: day 1, day 2, day 3, day 4, day 5, day6, day 9, day 12 and day 16. A total of nine days are significant in the principal component space for Individual 1.

6.8.6 Comparison of Clusters and Significant Days to Calendar Days

There are four main clusters from the loadings plot for Individual 1. Cluster one = day 1, day 2 and day 11 and these are represented by Friday, Sunday and Monday. Cluster 2 is represented by day 3 = Monday, day 5 = Wednesday, day 8 = Sunday, day 10 = Tuesday, day 12 = Thursday, day 14 = Saturday and day 15 = Sunday. Cluster 3 is represented day 7 = Saturday, day 9 = Monday and day 13 = Friday. The third cluster is represented by day 6 = Thursday and day 16 = Monday. Even though the days of measurement are not consecutive, but the clustering of the specific calendar that exhibit similar contact activity. There are nine significant days for Individual 1 and these are Friday, Sunday, Monday, Tuesday, Wednesday, Thursday, Monday, Thursday and Monday. The calendar days that are significant show similarities. A total of 9 days would be required at least to represent the contact activity of Individual 1

6.8.7 Summative Comparison of Class-4 Individuals

Table 6.7: Summary of Main Results for Class 4

Individual ID	Cumulative Variance	Groupings of Similar Days		Number of Clusters: Loadings Plot	Number of Outliers: Loadings Plot	Number of Significant Days
		Time Trends	Loadings Plot			
Individual 1	63%	{d1,d2,d3,d11,d12}; {d4,d5,d6,d9,d10}; {d7,d8,d13}; {d14,d15}	{d1,d2,d11}; {d3,d5,d8,d10,d12,d14,d15}; {d7,d9,d13}; {d6,d16}	4	1	9
Individual 5	49%	{d2,d7}; {d3,d4,d9}; {d6,d8,d15}; {d11,d12}	{d1,d10}; {d7,d8,d13}; {d11,d12}; {d3,d5,d6,d15}; {d4,d9}	5	2	9
Individual 6	67%	{d1,d2,d4,d7,d8,d14}; {d3,d5,d9,d10,d12,d12}	{d2,d3,d4,d5,d6,d7,d8,d9,d10,d12,d13,d14}	1	3	12
Individual 7	63%	{d1,d9,d10}; {d2,d3,d11}; {d4,d5,d6,d8,d14}; {d7,d12}	{d1,d15}; {d2,d4}; {d3,d5,d7,d11,d12,d14,d16}; {d9,d10,d13}	4	2	11
Individual 10	73%	{d1,d15}; {d2,d4,d9,d10,d11,d12}; {d3,d5,d6}	{d1,d8}; {d7,d13,d14,d15}; {d2,d3,d4,d5,d9,d10,d11,d12,d16}	2	1	9
Individual 15	66%	{d1,d3,d13}; {d4,d11}; {d6,d12,d14}; {d7.d9,d10}	{d1,d4,d11}; {d3,d5}; {d2,d6,d7,d8,d9,d10,d12,d13,d14}	3	0	10
Individual 17	58%	{d2,d4}; {d3,d7,d15}; {d9,d10}; {d13,d14}	{d7,d11}; {d2,d3,d14}; {d1,d5,d6,d8,d9,d10,d12,d13,d15}	3	1	8
Individual 19	71%	{d1,d12,d18}; {d2,d3}; {d6,d14}; {d7,d8,d11,d13,d15};	{d1,d12}; {d2,d5,d10}; {d12,d18}; {d3,d7,d16}; {d8,d6,d14}	5	4	10

This group of has the highest number of individuals and the number of logged contact days ranges from 14 to 18 days. This group of individuals also demonstrate a series of clustering across similar days of contact events. This group in particular, has the largest number of days forming a single cluster. For example, Individual 6 has one cluster of 12 similar days. Individual 10, 15 and 17 have a large cluster of 9 days. Individual 1 and 7 have a large cluster of 7 days. This shows that for these individuals, the higher the number of contact days, the greater the similarity between the days. It also shows that for these individuals, the contact activity is fairly consistent and similar across the days. The number of significant days in the principal component space ranges from 8 to 11. This means that for these individuals at least 8 days of measurement are required to represent the contact activity.

6.9 Summary

The main analysis presented in this chapter shows that for all individuals that have been grouped into the four Classes based on number of available days of contact activity, demonstrate a consistent clustering of similar days of contact activity. The first three principal components for all individuals represents a good level of variation in the original data and has shown to be a good choice for reducing the dimensionality of the original data set to three dimensions for analysis. Each individual has at least two clusters of similar days, with the exception of Individual 4 and Individual 6, from Class 2 and Class 4, who have a single cluster of five and twelve days of similar days respectively. There is also a similarity in some of the calendar days regarding contact activity. For an arbitrarily chosen threshold of 10%, the number of significant days in the principal components space ranges from 7-9 for Class 1; 5-7 for Class 2; 4-11 for Class 3 and 8-11 for Class 4.

The next chapter will critically discuss the results and use this to answer the research question, which is to determine whether the contact activity of an individual varies or remains similar across the days and the minimum number of days required adequately measuring and representing an individual's contact activity.

Chapter 7 – Discussion

This chapter will discuss the results presented in Chapter 6, in order to the answer the research question of this research which is: do contact events between individuals exhibit similarity across the days and the minimum number of days required to measure and represent an individual's contact activity. This chapter is structured as follows: section 7.1 will discuss the similarity of days. Section 7.2 will discuss the number of days required to represent an individual's contact activity. Section 7.3 will discuss the limitations of this research. Section 7.4 will discuss the impact of this research and section 7.5 will discuss future work recommendations.

7.1 Similarity of Days

Twenty datasets were analysed in this research, which represent the contact events over a period between 6 – 18 days of 20 individuals from a pilot study conducted by the author. The individuals were grouped into four categories based on the number of available days of contact activity. These categories were Class 1, representing three individuals who had 12 consecutive of contact events; Class 2, representing three individuals who had 7 days or less of contact activity; Class 3, representing six individuals with days of contact activity between 8 and 13; Class 4, representing eight individuals with 14 or more days of contact activity. The results of these individuals are summarised in Table 6.4, Table 6.5, Table 6.6 and Table 6.7, respectively.

All individuals demonstrate a clustering of days similar in contact events. These clusters are determined from an automated process then verified visually. Visual inspection of the time trends, show the contact events over a 24-hour period, across the days of the experiment, and the loadings plot, which show the span of the variables i.e. days of the experiment in the first three principal components space. Days that have similar patterns form a cluster. The more clustering is observed, the greater the variation in contact activity across the days. If an individual's contact activity is the same throughout the days of the experiment, then a single cluster would be expected. All the individuals have at least two clusters of similar days,

with the exception of Individual 4 from Class 2 and Individual 6 from Class 4, who have a single cluster of 5 days and 12 days, respectively. Individual 4 has 7 days of contact activity, of which five form a single cluster and two are outliers. Individual 6 has 12 days forming one cluster and three days are outliers. For these two individuals, a conclusion can be made about their contact activity being similar throughout the experiment period, with a few outliers.

The rest of the individuals in this analysis, have at least two or more clusters of similar days. This means that these individuals have at least two groupings of days of contact activity. This means that the contact activity is not the same throughout the experiment, but can be grouped according to clusters of similarity. The higher the number of clusters, the greater the variability in contact activity. Class one represents individual with 12 days of consecutive measurements. This group of individuals is important in demonstrating effectively the similarity or dissimilarity of contact events over time. This particular group shows that the calendar days that form similar clusters are far apart. The example from Individual 11 with two clusters shows that cluster-1 is formed by Saturday, Monday, Thursday and Saturday. The second cluster is formed by Friday, Sunday and Tuesday. This pattern of non-consecutive calendar days that form a cluster is consistent among all individuals. This result shows that contact events are not the same for all the days, even in the case of consecutive days of measurement.

Class- 2 shows individuals with fewer days of measurement ranging from 6-7 days. There are few clusters of similar days within this group as seen from Table 6.5 i.e. 1-2 clusters of similar days. From this class of individuals, the fewer the days of measurement, the smaller the number of clusters, which demonstrates less variation across the days of measurement. Even with these smaller numbers of clusters, the calendar days that form a cluster are still far apart or at least a day or two apart as shown in the analysis. Class 3 and Class 4, have more days of measurement. The results from Table 6.6 and Table 6.7 shows that there are more clusters of similar days formed from these group of individuals except for Individual 3, Individual 6, Individual 10 and Individual 12, who have 1-2 clusters. The rest of the individuals have 3-5 clusters of similar days. The results show that when increasing the number

of measurement days introduces more heterogeneity in contact events, hence the increase in number of clusters. However, from Class-4, the results show that in general, particularly Individual 1, Individual 5, Individual 7, Individual 10, Individual 15 and Individual 17 have large clusters i.e. more days forming a single cluster and this is the difference between Class-1 and Class-2, the number of days forming a single cluster. This means that increasing the number of measurement days to greater than 14 may introduces redundancy, as more days demonstrate similarity by forming a single cluster.

To answer the research question, the days of contact activity vary for a majority of the individuals and the number of clusters observed as shown in Table 6.4, Table 6.5, Table 6.6 and Table 6.7 represents this. Class 1, have a number of clusters that vary from 2 to 5; Class 2, excluding Individual 4, have two clusters of similar days; Class 3, have a number of clusters that vary from 2 to 4, and Class 4, excluding Individual 6, have a number of clusters that vary from 2 to 5. Therefore, the days of contact activity between individuals do exhibit variation over time and this is represented by the number of clusters of similar days.

7.2 Minimum Number of Days Required to Represent Contact Activity

As indicated in Chapter 3, studies of this kind have typically measured contact activity for 2-5 days. The average number of contact events per day is calculated, to determine if there is a variation in contact activity. An inference is made, based on this calculation that the contact activity remains consistent. A cosine similarity has also be calculated by *Kiti et al* and *Ozella et al* [19], [36], which takes into account the similarity of neighbours of a network node. This cosine similarity is given by the equation below obtained from [36]

$$sim(i) = \frac{\Sigma_j \left(w_{ij,1}\ w_{ij,2}\right)}{\sqrt{\Sigma_j (w_{ij,1}^2)}\ \sqrt{\Sigma_j (w_{ij,2}^2)}}$$

Where $w_{ij,1}$ and $w_{ij,2}$ are the weights (cumulative sum of contact events) on the link between node *i* and *j* measured between two time points (i.e. two consecutive days), respectively [19], [36]. The cosine similarity is 1 if individual *i* had the exact same neighbours on the network for the two days of measurement, with the same fraction of time and 0 for the opposite [19], [36]. This is a measure of how similar the interactions between individuals in a household are. However, it does not account for the variation of contact events across the number of days for a given individual, because the measurement period was only 2-5 days, of which only two days of data was used for analysis. The limitation with this approach is that a household has a finite number of members and there is a high possibility that these members will be in contact with each other, most of the time. Therefore, the cosine similarity and average number of daily contacts, does not fully represent the temporal changes in an individual's contact activity, especially over an extended period.

Similarly, work done by *Kiti et al*, in quantifying contact patterns in a rural community in Kenya, measured contact events for three consecutive days [19]. Work done by *Vanhems et al* [21] and *Isella et al* [23], measured contact patterns in a hospital ward and paediatric ward for 4 days and 5 days respectively. *Stehle et al* [7] measured contact patterns in a primary school for a period of 2 days. All of these studies used wearable proximity sensors. All of these studies have a potential of misrepresenting the contact activity of the individuals in their respective study, given the variation of contact activity across the days, as the results shows in this research.

Principal Component Analysis is used in this research to determine the variation in contact activity and find clusters of similar days. The contribution of the original variables i.e. the days is determined by the loadings. Each loading determines the correlation between the original variable and the principal component. In this research, the first three principal components were used to retain a good representation of variation in the data. These three principal components were used to visualise the loadings representing the days of the experiment. A threshold value is typically specified to select the variables that have a significant contribution to

the principal components. This threshold depends on the type of data and type of analysis required. Some studies would choose variables that have 60% contribution to the principal components, other studies would choose a threshold value equal to (1 / number of variables) for uniform variables e.g. if there are 12 variables, a threshold value of 1/12 = 0.083 would be used to retain variables with a contribution of 8.3% and above [50]. In this research, to demonstrate the significance of feature selection, an illustration was made with a threshold of 10% to show that some variables are insignificant to the principal components and such variables can be left out.

The results in the analysis show that for all the individuals, the number of days that have at least a 10% contribution to the first three principal components ranges from 5-11. Days that have a significant contribution to the principal components, are the days that contain the most variation in the data, which means they contain the most important information towards the new set of variables i.e. principal components. Therefore, the number of significant days in the principal components space, would represent the minimum number of days to measure an individual's contact activity. This is subject to the choice of the threshold level used to retain significant days and this value will have to be determined experimentally and validated in a bigger cohort of individuals.

7.3 Limitations of Research

There are several limitations to this research that may affect the conclusions made from the results.

7.3.1 Quality of Data

The first and important limitation of this research is the quality of the data collected. The data collected for this research was from ten household deployed in two sites, a rural and a peri-urban area. Five households were selected in both areas; this was constrained to ethics requirements, because this was a pilot study. Fifty-one sensors were deployed to 51 participants, from which 45 sensors were returned. From the returned sensors, only 20 sensors had usable data of at least 5 days of measurement.

This reduced the sample size significantly and made it very difficult to carry out household level analysis of contact events, because there was only one household of complete participant data. The percentage usable data as shown from Figure 6.3, demonstrates that 60% of the participants i.e. 12/20 have 60% of clean contact data and the remaining proportion of individuals had ≤ 50% of clean data. This means that a significant portion of contact events are false contact events.

The compliance of participants to the experiment was also a huge limiting factor. This entire experiment hinges on the compliance of individuals to wear the proximity sensor as required. Failure to do so leads to a lack of logged contact activity and a fewer number of consecutive days of measurement as seen in this research. The days of no logged contact activity had to be removed from the analysis and this has a potential of misrepresenting contact events and the patterns thereafter.

7.3.2 Representation of True Contact Events

The contact events that are logged are limited to the individuals being in close-range physical proximity of <1.5 m to log an event. In some cases, individuals are not necessarily in that close range proximity to be in contact and/or are not wearing the sensors as discussed in the preceding section. This leads to a number of false negatives, which are contact events that should have been logged, but were missed. Unfortunately, there is no way of recovering these false negatives, unless supplementary data like contact diaries, user-filled time logs of contact events, are implemented. False negatives have a potential risk of misrepresenting the variation or similarity of contact events over time.

Secondly and more importantly, individuals' logged contact events are inter-related to the logged contact events of other members of the household. The contact events of each individual, depends on the presence of another individual wearing the sensor. This poses a question of how true are the individuals contact events, if they depend on other individuals to log contact events. Can we truly rely solely on wearable proximity sensors to find patterns of contact activity, if compliance is an important measure and often lacks in studies of this kind? Specifically, when

measuring the variation of contact activity over time, do individuals exhibit this variation of contact activity over time or is it a function of other members being present and wearing the sensors. This is an important limitation in the representation of individual's true contact activity.

The cleaning of the data relied solely on the status flag value indicating motion of the sensors. Although this is important in determining whether the proximity sensors are worn or not, there are cases where individuals could be sitting still for some time and a status flag value, based on changes in accelerometer data, might log that state as stationary, therefore those true contacts might be cleaned out as false contacts. Secondly, the sensors need to be directly in front of each other to log contact events. In cases where individuals are sitting adjacent to each other e.g. watching the TV from the same angle or walking adjacent to each other, those contact events may not be logged. This represents once again, the false negatives mentioned above and it is a known challenge in studies of this kind. Batteries also run out of power and this was the case in the pilot study conducted by the author. Batteries need to be consistently replaced, when deploying a study of this kind for a long period. Failing to do so, leads to missed contact events and a misrepresentation of an individual's contact activity.

7.3.3 Limitations of PCA

PCA requires a dataset to be normalized and standardised in order for each variable to contain the same weight, otherwise this leads to bias and certain variables will dominate the analysis [50]. In this research, the days that had no contact activity were removed, because there is no information. However, the days that had a single contact event, would dominate the principal component analysis, because of the weight they carried and these days were removed as shown in Table 6.2, the calendar grid that maps the variables to the calendar days. Although this data preparation step is necessary for PCA, removing those days of single contact activity, might misrepresent the individual's contact patterns. It could also be that on those particular days there was little or no compliance from other members of the household, leading to these very sparse contact events on those days. In this research, PCA has been shown to find good patterns in contact activity, through the

clustering of similar days and determining the minimum number of days required for measuring contact events. This is subjected to good and consistent data of contact activity.

7.4 Impact of Research

Contact patterns between individuals are important in complex behavioural science and network science often used in epidemiological studies. Contact patterns between people have an important association in assisting the understanding of transmission of infectious diseases. Contact patterns are used to parameterize transmission models in public health research, determine contact rates between people, age-mixing patterns between people, internet of things and social network analysis. This research specifically, will contribute towards understanding the temporal patterns of individual's contact activity. Understanding that there is clustering in individual's contact activity, will help develop better models of complex behavioural science.

A major contribution of this research is the recommendation of the minimum number of days required for representing an individual's contact activity. This shows that previous studies may have potentially misrepresented the contact patterns between individuals by only measuring for a shorter period. This will be important in shaping future studies that will use wearable proximity sensors to measure and represent contact patterns between individuals.

The cleaning of algorithm proposed in this research will enable a more efficient and robust cleaning of contact events that can be scaled up to datasets of any size. The representation of temporal contact events in a matrix format, will allow for automated analysis techniques in signal processing, image processing, data science etc. to find new ways of representing and analysing data of this kind.

7.5 Future Work

Future work will need to apply the techniques used in this research in a much bigger dataset in order to compare and validate the findings. A compliance measure of

other individuals in a household, needs to be factored into the representation of an individual's true contact events over time. This will require complete household data from all participants to cross-validate logged contact events.

Future studies could also look into a comparison of the number of clusters versus a 10% threshold (or threshold of any kind) in representing the minimum number of days required for measurement. Each cluster represents the number of similar days in the principal component analysis. If these days are similar, this can be considered as a representation of a single day's worth of contact activity. The reason this was not explored in this research, is that every individual had outliers that did not form a cluster and the weight of these outliers needs to be determined in order to conclude that each cluster represents a day's measurement of contact activity. However, a specified threshold to retain the number of important days remains a valid method. Future studies could also explore what the minimum number of days required is for a good representation of an individual's contact activity. This research unfortunately was limited by the poor quality of data collected and the findings may not be generalizable, but rather a reflection of the artefacts in the data.

The cleaning algorithm needs to be applied to a secondary dataset that has been cleaned using a different method. In this way, the derived cleaning algorithm can be validated using a secondary and bigger dataset. This is open to future work and could not be done using the current cohort of individual datasets, because of the quality of the data.

Classification of contact activity across the days can also be done on this type of data. Individuals can be classified according to the contact activity and asses the association of such classes of contacts to the spread of infectious diseases.

Public health studies looking into household contacts can use this type of data to classify different households according to the individual contact activity and this may or not have a significance in assessing factors that may drive household transmission of infectious diseases.

7.6 Summary

In this chapter, a discussion of the results is presented to answer the research question in this masters. It is found that there are similarities between some of the days of an individual's contact activity and some days are different. This is represented by the clustering of similar days. The more clusters an individual has, the greater the variation in the contact activity over the days. The number of days required to measure an individual's contact activity, could not be conclusively determined, because this depends on the threshold chosen for feature selection and that the number of significant days vary across all individuals from 5-11 days. However, it was demonstrated in this research that 2-5 days of measurement would not be sufficient to represent an individual's contact activity, because of the clustering of days that are observed and that the calendar days that form the clusters are not consecutive calendar days. Research limitations of the quality of data, the representation of true contact events and limitations of PCA in this study, were discussed and recommendations on potential future work were made.

Chapter 8 – Conclusion

Representing wearable proximity sensor data in a manner that enables automatic analysis of the data will play an important role in expanding the way this type of data is understood and processed. Proximity sensor data plays an important role in the understanding, measuring and modelling of human social behaviour and how this plays a role in complex social systems. Studies that used wearable proximity sensors, including earlier methods of contact diaries and friendship surveys, have made assumptions on the similarity of contact patterns over time. These assumptions have led to shorter periods of measurements of contact events, which range from 2-5 days. Many of the social behavioural models, including those that are used in public health for modelling the spread of infectious diseases, have parameterised models based on data from previous studies that only measured contact patterns for a short period of time. The aim of this research was to investigate and quantify the similarity of contact events over time for individuals and determine the minimum number of days required, to adequately represent an individual's contact activity.

An efficient cleaning algorithm was implemented that used both the contact and status dataset, represented as matrices of daily logged contact events and status flag values. Using these two datasets in a matrix representation, an element-wise multiplication of the contact data and status data, made it easier to clean the data, by using the logged movement status of the sensor. This algorithm is efficient, easy to implement and can be scaled to larger datasets, without requiring individual inspection of each sensor to determine what a false and true contact is.

Individuals were categorised in four classes, based on the number of available days of contact events. Class-1 represents individuals with consecutive days of contact events. Class-2 represents individuals with seven days or less of available contact events. Classs-3 represents individuals with available days of contact events between 8-13. Class-4 represents individuals with contact events of 14 days or more. This categorisation made it easier to group individuals and compare the similarity of days within each class and across the individuals.

From visual inspection of the data, all the individuals regardless of the category, exhibited similarity and variation in contact events across the days. The similarity was represented by clusters of similar days identified and the variation was represented by the total number of clusters of days for each individual. The higher the number of clusters, the more variation in contact activity. Principal component analysis was used to reproduce the findings from the visual inspection of contact events in a quantitative manner. All the individuals had at least two or more clusters of similar days, except for Individual 4 from Class 2 and Individual 6 from Class 4. These individuals had one main cluster of similar days with a greater proportion of days forming the cluster. The analysis of loadings plots which shows the clustering of the different days, shows that some days are similar and other are different. For studies that measured contact events for 2-5 days, will most likely under represent the contact activity of individuals, because of the clustering of days.

The second part of the research was aimed at determining the minimum number of days required to adequately represent an individual's contact activity. PCA allows for the representation of the original inter-correlated variables in a new set of uncorrelated variables called principal components. Each principal component is a linear combination of the original variables. The contribution of each original variable to the principal components is determined the value of the loadings. It was demonstrated in this research that some variables are more significant in the principal component space than others. The smaller the value of the contribution, the more insignificant that variable is to the principal component. There has been no experimental evidence on data of this kind to suggest a threshold of retaining variables that are more significant and therefore, should be retained. A threshold of 10% was used in this research to show how some variables can be discarded from the analysis. With this threshold, the minimum number of variables that are significant in the principal component space varied across all individuals from 5 - 11 days. This means that at least five days of measurement for some individuals is required to adequately represent the contact activity of those individuals. The data set was too small; to determine experimentally what this ideal threshold should be. This should be explored in future studies with a bigger cohort.

The findings of this research will be important in transmission models that use contact patterns to calibrate their models for simulation of disease transmission. This will also play an important role in modelling human behaviour in complex social behavioural systems. The quantitative representation of the proximity sensor data, allows for many other analysis tools in signal processing, machine learning and big data analytics to mention a few, to broaden the scope for representation and analysis of data of this kind.

Future work should look into applying the methods used in this research to a bigger cohort of individuals to validate and compare the findings. A compliance measure should be considered to factor in the logged contact events of other members in the household to cross-validate the contact events for each individual in the household. A comparison of the number of clusters versus a threshold for retaining significant variables, should be explored as a means to determining the minimum number of days required to represent an individual's contact activity. Machine learning tools can be considered in future studies towards the classification of individual contact activity and assessing the association of these classes of individual contact activities, in assisting the understanding of transmission dynamics of infectious diseases, required in public health research.

References

[1] C. Cattuto, W. Broeck, A. Barrat, V. Colizza, J. F. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed RFID sensor networks," *PLoS One*, vol. 5, 2010.

[2] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, and J. E. Larsen, "Measuring large-scale social networks with high resolution," *PLoS One*, vol. 9, 2014.

[3] M. Atzmueller, L. Thiele, G. Stumme, and S. Kauffeld, "Analyzing group interaction on networks of face-to-face proximity using wearable sensors," in *2018 IEEE International Conference on Future IoT Technologies (Future IoT)*, 2018, pp. 1–10.

[4] M. Starnini, B. Lepri, A. Baronchelli, A. Barrat, C. Cattuto, and R. Pastor-Satorras, "Robust modeling of human contact networks across different scales and proximity-sensing techniques," *ArXiv170706632 Phys.*, Jul. 2017.

[5] M. C. Kiti, T. M. Kinyanjui, D. C. Koech, P. K. Munywoki, G. F. Medley, and D. J. Nokes, "Quantifying Age-Related Rates of Social Contact Using Diaries in a Rural Coastal Population of Kenya," *PLOS ONE*, vol. 9, no. 8, p. e104786, Aug. 2014.

[6] T. Smieszek, S. Castell, A. Barrat, C. Cattuto, P. J. White, and G. Krause, "Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes," *BMC Infect. Dis.*, vol. 16, no. 1, p. 341, Jul. 2016.

[7] J. Stehlé *et al.*, "High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School," *PLOS ONE*, vol. 6, no. 8, p. e23176, Aug. 2011.

[8] N. Eagle, A. (Sandy) Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Natl. Acad. Sci.*, vol. 106, no. 36, pp. 15274–15278, Sep. 2009.

[9] W. J. Edmunds, C. J. O'Callaghan, and D. J. Nokes, "Who Mixes with Whom? A Method to Determine the Contact Patterns of Adults That May Lead to the Spread of Airborne Infections," *Proc. Biol. Sci.*, vol. 264, no. 1384, pp. 949–957, 1997.

[10] J. M. Read *et al.*, "Social mixing patterns in rural and urban areas of southern China," *Proc. Biol. Sci.*, vol. 281, no. 1785, p. 20140268, Jun. 2014.

[11] N. Hens, N. Goeyvaerts, M. Aerts, Z. Shkedy, P. Van Damme, and P. Beutels, "Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium," *BMC Infect. Dis.*, vol. 9, p. 5, Jan. 2009.

[12] J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, and R. Mikolajczyk, "Social contacts and mixing patterns relevant to the spread of infectious diseases," *PLoS Med*, vol. 5, 2008.

[13] P. Horby *et al.*, "Social Contact Patterns in Vietnam and Implications for the Control of Infectious Diseases," *PLOS ONE*, vol. 6, no. 2, p. e16965, Feb. 2011.

[14] M. L. Stein *et al.*, "Comparison of Contact Patterns Relevant for Transmission of Respiratory Pathogens in Thailand and the Netherlands Using Respondent-Driven Sampling," *PLOS ONE*, vol. 9, no. 11, p. e113711, Nov. 2014.

[15] A. J. Conlan, K. T. Eames, J. A. Gage, J. C. Kirchbach, J. V. Ross, and R. A. Saenz, “Measuring social networks in British primary schools through scientific engagement,” *Proc Biol Sci*, vol. 278, 2011.
[16] D. J. Toth, M. Leecaster, W. B. Pettey, A. V. Gundlapalli, H. Gao, and J. J. Rainey, “The role of heterogeneity in contact timing and duration in network models of influenza spread in schools,” *J R Soc Interface*, vol. 12, 2015.
[17] N. Eagle and A. (Sandy) Pentland, “Reality Mining: Sensing Complex Social Systems,” *Pers. Ubiquitous Comput*, vol. 10, no. 4, pp. 255–268, Mar. 2006.
[18] SocioPatterns Initiative, “SocioPatterns.” [Online]. Available: http://www.sociopatterns.org/. [Accessed: 19-Mar-2018].
[19] M. C. Kiti *et al.*, “Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors,” *EPJ Data Sci.*, vol. 5, no. 1, p. 21, Dec. 2016.
[20] J. Fournet and A. Barrat, “Contact patterns among high school students,” *PLoS One*, vol. 9, 2014.
[21] P. Vanhems *et al.*, “Estimating Potential Infection Transmission Routes in Hospital Wards Using Wearable Proximity Sensors,” *PLOS ONE*, vol. 8, no. 9, p. e73970, Sep. 2013.
[22] N. Voirin, C. Payet, A. Barrat, C. Cattuto, N. Khanafer, and C. Regis, “Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital,” *Infect Control Hosp Epidemiol*, vol. 36, 2015.
[23] L. Isella *et al.*, “Close Encounters in a Pediatric Ward: Measuring Face-to-Face Proximity and Mixing Patterns with Wearable Sensors,” *PLOS ONE*, vol. 6, no. 2, p. e17144, Feb. 2011.
[24] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, “A high-resolution human contact network for infectious disease transmission,” *Proc Natl Acad Sci U A*, vol. 107, 2010.
[25] A.-L. Barabasi, “Network Science by Albert-László Barabási.” [Online]. Available: http://networksciencebook.com/. [Accessed: 19-Mar-2018].
[26] M. J. Keeling and K. T. D. Eames, “Networks and epidemic models,” *J. R. Soc. Interface*, vol. 2, no. 4, pp. 295–307, Sep. 2005.
[27] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.
[28] C. Cohen, “Prospective Household observational cohort study of Influenza, Respiratory Syncytial virus and other respiratory pathogens community burden and Transmission dynamics in South Africa - The PHIRST Study,” *Study Protocol*, Jul. 2018.
[29] Bitmanufactory, “Proximity Sensors for the Social Web,” *OpenBeacon*. [Online]. Available: https://www.bitmanufactory.com/. [Accessed: 19-Mar-2018].
[30] M. Fullen, P. Schuller, and O. Niggemann, “Defining and validating similarity measures for industrial alarm flood analysis,” in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Emden, 2017, pp. 781–786.
[31] K. Ahmed, I. Izadi, T. Chen, D. Joe, and T. Burton, “Similarity Analysis of Industrial Alarm Flood Data,” *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 2, pp. 452–457, Apr. 2013.

[32] Fan Yang, S. L. Shah, and Deyun Xiao, “Correlation analysis of alarm data and alarm limit design for industrial processes,” in *Proceedings of the 2010 American Control Conference*, Baltimore, MD, 2010, pp. 5850–5855.
[33] A. W. Al-Dabbagh, W. Hu, S. Lai, T. Chen, and S. L. Shah, “Better understanding of process operation using performance metrics and visualization plots,” in *2017 IEEE Conference on Control Technology and Applications (CCTA)*, 2017, pp. 2015–2020.
[34] T. D. Butters, S. Guttel, J. L. Shapiro, and T. J. Sharpe, “Statistical cluster analysis and visualisation for alarm management configuration,” in *Asset Management Conference 2014*, 2014, pp. 1–6.
[35] A. Griparis, D. Faur, and M. Datcu, “Feature space dimensionality reduction for the optimization of visualization methods,” in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 1120–1123.
[36] L. Ozella *et al.*, “Close encounters between infants and household members measured through wearable proximity sensors,” *PLOS ONE*, vol. 13, no. 6, p. e0198733, Jun. 2018.
[37] R. E. Ziemer, W. H. Tranter, and D. R. Fannin, *Signals and Systems: Continuous and Discrete*, 4 edition. Upper Saddle River, NJ: Pearson, 1998.
[38] “Introduction to Discrete-time Signals and Systems | Robert I. Damper | Springer.” [Online]. Available: https://www.springer.com/gp/book/9780412476501. [Accessed: 14-Nov-2018].
[39] B. Scholz-Reiter, C. Toonen, J. T. Tervo, and D. Lappe, “Sampling of Variables in Discrete-Event Simulation using the Example of Inventory Evolutions in Job-Shop-Systems Based on Deterministic and Non-Deterministic Data,” vol. 5, no. 1, p. 6, 2011.
[40] G. Zhao, “The Property of Kippenhahn Curves of Hadamard Products and Applications,” in *2009 International Conference on Computer Technology and Development*, Kota Kinabalu, Malaysia, 2009, pp. 568–571.
[41] L. Feilong and H. Xiqin, “Lower bound for the eigenvalue of Hadamard product of an M-matrix and its inverse,” in *2011 Chinese Control and Decision Conference (CCDC)*, Mianyang, China, 2011, pp. 3677–3680.
[42] P. Poekaew and P. Champrasert, “Adaptive-PCA: An event-based data aggregation using principal component analysis for WSNs,” in *2015 International Conference on Smart Sensors and Application (ICSSA)*, Kuala Lumpur, Malaysia, 2015, pp. 50–55.
[43] V. Sathish, S. D. Sudarsan, and S. Ramaswamy, “Event Based Robot Prognostics Using Principal Component Analysis,” in *2014 IEEE International Symposium on Software Reliability Engineering Workshops*, Naples, Italy, 2014, pp. 14–17.
[44] A. Onishi and K. Natsume, “Ensemble regularized linear discriminant analysis classifier for P300-based brain-computer interface,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, 2013, pp. 4231–4234.
[45] J. Shlens, “A Tutorial on Principal Component Analysis,” *Int. J. Remote Sens.*, Apr. 2014.

[46] Y. Wen and P. Shi, “Image PCA: A New Approach for Face Recognition,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07*, Honolulu, HI, 2007, pp. I-1241-I–1244.
[47] T. Jagadesh and R. J. Rani, “A novel speckle noise reduction in biomedical images using PCA and wavelet transform,” in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2016, pp. 1335–1340.
[48] W. Liao *et al.*, “Hyperspectral image deblurring with PCA and total variation,” in *2013 5th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Gainesville, FL, USA, 2013, pp. 1–4.
[49] Z. Lou, J. Tuo, and Y. Wang, “Two-step principal component analysis for dynamic processes,” in *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*, Taipei, Taiwan, 2017, pp. 73–77.
[50] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.
[51] H. Hasan and N. M. Tahir, “Feature selection of breast cancer based on Principal Component Analysis,” in *2010 6th International Colloquium on Signal Processing its Applications*, 2010, pp. 1–4.
[52] Chioua M, Bauer M, Chen S, Schlake J, Sand G, Schmidt W, Thornhill N, “Plant-wide root cause identification using plant key performance indicators (KPIs) with application to a paper machine,” *Control Eng. Pract.*, vol. 49, pp. 149–158, Apr. 2016.
[53] Y. Yeh, C. Chen, C. W. Chiou, and T. Chu, “A reliable feature selection algorithm for determining heartbeat case using weighted principal component analysis,” in *2016 International Conference on System Science and Engineering (ICSSE)*, 2016, pp. 1–4.
[54] F. Song, Z. Guo, and D. Mei, “Feature Selection Using Principal Component Analysis,” in *Engineering Design and Manufacturing Informatization 2010 International Conference on System Science*, 2010, vol. 1, pp. 27–30.

Appendix A – Ethics Clearance

University
of the Witwatersrand,
Johannesburg

Human Research Ethics Committee: (Medical)
FWA Registered No IRB 00001223

SECRETARIAT: Suite 189, Private Bag x2600, Houghton 2041, South Africa Tel: +27-11-274 9200 Fax: +27-11-274 9281

Prof C Cohen,

Centre Head: Epidemiology
Centre for Respiratory Diseases & Meningitis
National Institute for Communicable Disease
a division of the National Health Laboratory
1 Modderfontein Rd Sandringham
2193

Fax: 011 882 9979

EMAILED

27 July 2018

Dear Prof Cohen,

PROTOCOL: PHIRST - A PROSPECTIVE HOUSEHOLD OBSERVATIONAL COHORT STUDY OF INFLUENZA, RESPIRATORY SYNCYTIAL VIRUS AND OTHER RESPIRATORY PATHOGENS COMMUNITY BURDEN AND TRANSMISSION DYNAMICS IN SOUTH AFRICA (THE PHIRST STUDY)

ETHICS REFERENCE NO: 150808

RE : APPROVAL FOR PROTOCOL AMENDMENT

We acknowledge receipt of your letter dated 13 July 2018 with the following documentation pertaining to the above-captioned trial.

Amendment Date:	13-Jul-2018	Amendment Version:	Protocol Amendment
Amendment Number:		Received Date:	13-Jul-2018

The following has been approved by the Wits Human Research Ethics Committee: (Medical):

* Protocol: Community burden and transmission of influenza and RSV (PHIRST study) Amendment dated 13 July 2018 (Tracked)
* Summary of Changes

Ethics Approval Date: 27 July 2018

The University of the Witwatersrand, Human Research Ethics Committee Approval Granted for the above mentioned study is valid for five years. Where required by Sponsor to have approval on a more frequent basis it remains the responsibility of the Sponsor and Investigator to apply for continuing review and approval, or for the duration of the Trial.

1. THIS APPROVAL IS SUBJECT TO THE FOLLOWING PROVISOS:

* A copy of the SAHPRA Approval and/or SAHPRA Notification letter must be submitted to the Ethics Regulatory Office Secretariat before the study commences / or where an Amendment may be implemented (IF SAHPRA APPROVAL / NOTIFICATION IS APPLICABLE). It remains the responsibility of the Principal Investigator and/or Sponsor to ensure that the relevant approvals are in place.

* The study is conducted according to the protocol submitted to the University of the Witwatersrand, Human Research Ethics Committee. Any amendments to the protocol must first be submitted to the Human Research Ethics Committee for approval.

* During the study, the University of the Witwatersrand, Human Research Ethics Committee is informed immediately of :
  - Any Unexpected Serious Adverse Events or Unexpected Adverse Drug Reactions, which, in the Investigator and/or the Sponsor's opinion are suspected to be related to the study drug. (Refer to POL-IEC–001 and SOP-IEC–005, Item 3.4).
  - Any data received during the trial which, may cast doubt on the validity of the continuation of the study .

* The University of the Witwatersrand, Human Research Ethics Committee is notified of any decision to discontinue the study and the reason stated.

* The Investigators authorised by this approval participate in this study. Additional Investigators shall be submitted to the University of the Witwatersrand, Human Research Ethics Committee for approval prior to their participation in the study.

* In the event of an authorised Investigator ceasing to participate in the study, the University of the Witwatersrand, Human Research Ethics Committee must be informed and the reason for such cessation given.

2. PRINCIPLES OF INFORMED CONSENT:

* The University of the Witwatersrand, Human Research Ethics Committee requires that in all studies, the Principles of Informed Consent are adhered to. This applies to volunteers as well as patients.

3. PROGRESS REPORTS:

* The University of the Witwatersrand, Human Research Ethics Committee requests that the MCC Progress Reports be submitted twice a year either in March and September or six monthly from start of study to the HREC Secretariat Office - 011 274 9281 and a report of the final results, at the conclusion of the study. (IF APPLICABLE)

4. REIMBURSEMENT TO PATIENTS FOR TRANSPORT:

* The Human Research Ethics Committee: (Medical) is in agreement that reimbursement per visit is according to the South African Health Products Regulatory Authority and that reimbursement should be appropriate according to the situation.

5. TRANSPORT AND STORAGE OF BLOOD AND TISSUE SAMPLES IN SOUTH AFRICA:

* If blood specimens are to be stored for future analysis and is planned that such analysis will be done outside Wits, then the blood must be stored at a facility in South Africa agreed with the relevant IRB, with release of sub-samples only once projects have been approved by the local Research Ethics Committee applicable to where the analysis will be done as well as by the Wits Human Research Ethics Committee: (Medical).

6. GENETIC TESTING

* The Human Research Ethics Committee: Medical; will not approve open-ended genetic testing as this does not fit the Human Research Ethics Committee criteria.

7. GOOD CLINICAL PRACTICE

The South African Department of Health, South African Health Products Regulatory Authority requires Good Clinical Practice (GCP) Training for all Investigators in Clinical Trials, and that GCP training be renewed every three (3) years.

As yet, there are no National Guidelines for the content of GCP courses. Until these are available the Wits Human Research Ethics Committee (Medical) will note courses completed by Investigators without approval of the content of the individual courses.

The above has been noted for the Ethics Committee information and records.

KINDLY FORWARD TO THE RELEVANT INVESTIGATORS / CRA / STUDY CO-ORDINATORS

Regards,

PROF CLEMENT PENNY
For and on behalf of the Human Research Ethics Committee: (Medical)

Appendix B – Proximity Pilot Study Design

B1. Introduction

In this section, a summary of all the operational procedures and design considerations are presented for the deployment of the proximity pilot study. The study was deployed in Agincourt, Mpumalanga and Jouberton, Northwest for a full three weeks in October 2017.

B2. PHIRST Study

This research is nested in the PHIRST Study (Prospective Household observational cohort study of Influenza, Respiratory Syncytial virus and other respiratory pathogens community burden and Transmission dynamics in South Africa). The PHIRST Study is run by the National Institute for Communicable Diseases (NICD) and is a household-level community cohort study implemented in two sites in South Africa. The first site is a rural community in the Mpumalanga Province, where the study is conducted within the Health and socio-Demographic Surveillance Site (HDSS) at the Medical Research Council (MRC)/ Wits Rural Public Health and Health Transition Unit (in Agincourt). The second site is a semi-urban community in the Jourberton Township in Klerksdorp, in the North West Province, under the Perinatal HIV Research Unit (PHRU) of the Chris Baragwanath Hospital. This study has been running for three years, from 2016-2018 and will conclude in October 2018. The data for this research is obtained from a study ran by the author within the PHIRST study, where ten households were approached and enrolled and individuals were asked to wear the proximity sensors for three weeks. The aim is to use these datasets to automatically identify contact patterns between individuals at a household level.

B3. Household and Participant Selection

A household is defined as people living in the same yard and share the same kitchen. People of the same household may live in different buildings on the same property i.e. a backyard room/building for sleeping purposes. People living in outside dwelling who do not eat communally with the main household are not considered household members even if they are family. For the purpose of this pilot, members who are not enrolled in the PHIRTS study will not be approached. This will ensure compliance as PHIRST participants are already familiar the process and the field workers have established a very close working relationship with the participants. A household must have at least 3 members and more than 80% of household members must consent to participate prior to the household being enrolled.

Randomization and Household Selection Criteria

A list of randomly selected households is generated at the beginning of the study at each site. The randomization list is generated by NICD for both sites (Klerksdorp and Agincourt). The randomization process is as follows:

1. A total of 15 households will be selected at random for each site to allow for refusals i.e. 15 households will be randomly selected from 55 households at Klerksdorp and 15 households will be randomly selected from 53 households at Agincourt.
2. Once the 15 households have been selected, the top 5 households that meet the criteria detailed below (point 3) will be selected and approached for enrolment.
3. The selection criteria will be as follows:
    a. One household will need to have at least one infant below the age of 1.
    b. One household will need to have at least one child between 1 and 7 years of age.
    c. One household will need to have at least one child (between 7 and 13)
    d. One household will need to have at least one child (between 13 and 19)
    e. One household will need to have at least one adult (above 20 years)

    The purpose of this criteria is to stratify the sample of households and participants according to the give age groups list above.
4. Once the top 5 households have been selected, each household will be approached to explain the study and sign the consent forms once the members have indicated an interest to participate.

5. If a household refuses to participate, the next eligible household which falls into the relevant subgroup will be approached.

B4. Consent Forms and Enrolment

There are four forms that the field workers go through with the participants for consent. It is important that the field worker go through the forms carefully with the participants before they can grant consent to participate in the study. These forms are:

- Assent Leaflet for Individuals 7 – 17 Years: This is an information leaflet for children between the age of 7 and 17. The purpose of this document is to explain the study clearly to the children, the requirements, their rights and all legal considerations.
- Pilot Study Informed Consent Form - Adults: This is the consent form for adults, explaining to the participant what it is they are consenting to.
- Pilot Study Informed Consent Form - Minor: This is the consent form for minors signed by the parent on behalf of the minor.

Once the consent forms have been signed and collected, the participants are now enrolled in the pilot study. For the pilot, the participants' details will be populated in an excel file where meta data will be created containing the participants' study ID, age, gender and name. This file will be used to create an inventory of all the proximity tags and associate these to the participants.

B5. Deployment of the Proximity Study

The deployment of the proximity study includes the packaging of the proximity sensors, the design of the pouch that will be worn by the participants and the daily operational procedures during the study.

Packaging of the proximity tags: The proximity sensor is a light-weight, coin shaped tag that contains sensitive electronic components on its surface. These components can be easily damaged and need to be protected appropriately. The following is done to ensure the safety of the users (participants) and the tag itself

- The tag is covered with a transparent plastic material, as show in Figure B.1, to protect against water, sweat and damage from possible scratches.
- The tag is safely placed in a small, square pouch made from fabric. This is both for aesthetics and protection.

- The tag is worn as a lanyard as shown in Figure B.2. The tag is made of a white material, this is done to make it easier for the school children and working adults to wear it underneath their clothes.
- A safety pin is sewn on the tag itself for infants and attached directly onto their clothes. This is done to protect the infants from possible strangulation from the cord of the lanyard.

Figure B.1*: Proximity tags covered in plastic for protection*

Figure B.2*: Front and Back of proximity tag pouch worn by the participants*

Daily Operational Procedures: The following are the procedures that the participants follow throughout the study and these are monitored by two weekly visits conducted by the field workers to monitor the participants and address any concern that arise:

- The participants are each given a white pouch containing the proximity tags, with their study ID's written on the back of the pouch. They are required to wear the tags every day, throughout the day for the duration of the study.
- Permission is granted by head masters of the schools, to allow the school-going children to wear the tags at school.
- Working adults are asked to wear the tags to work and beneath their clothes (if they wish).
- The tag is designed as a lanyard and can be worn even in bed time, with a very light material for comfort and convenience.
- The participants are encouraged to wear the tag at night. However, if they are uncomfortable they need to store them securely in the storage pouches which are provided as indicated in the below point and remember to put wear them in the morning.

- Storage pouches made of foil are provided and this will shield the radio frequency signals when the tags are not used.
- The storage pouches have to be stored in the same location, preferably the kitchen where they will be easily accessible and visible to remind the participants in case they forget to wear the tags.
- School children will wear the tags in the morning as they get ready for school and keep them on throughout the day.
- School children will need to keep the tags on after changing their school uniform. The head of the household or care giver will be tasked to monitor and ensure that the school children keep the tags when they come back from school and after changing their uniform.
- Care givers will need to clip the tag on to the clothes of the infants using the safety pin on the pouch. This is a protective measure as the lanyard cord may pose a risk to infants (i.e. strangulation)

Supplementary Data: This is the data that will be used to supplement the measurements from the tags and will include a time log, a daily visit questionnaire done by the field workers and an acceptability questionnaire at the end of the study.

1. Routine Visits

The field workers have a routine visit questionnaire, administered online using RedCap database platform. The questions address what the daily experience of wearing the tag is, if the participants forgot to where the tags and why, challenges encountered so far. This type of routine monitoring is to ensure that any problems that arise are dealt with timeously so that the data collected from the sensors can be as accurate and meaningful as possible.

B6. Data Quality Control Measures

There are a number of things that can go wrong with data collection, from missing data, damaged tags, swapping of tags etc. Some are beyond control, but there are measures that can be put in place to ensure that data collected is as good as possible.

1. Corrupted Data

Data can be easily corrupted by swapping tags of participants which will give an incorrect representation of the contact patterns. This can be done by either the participants themselves when they incorrectly wear someone else's tag or the field worker could assign a tag to the wrong participant. To avert this, the tags themselves are labelled at the back with the participants' study ID (shown in Figure B.3) and on the pouch itself, with the participants' study ID.

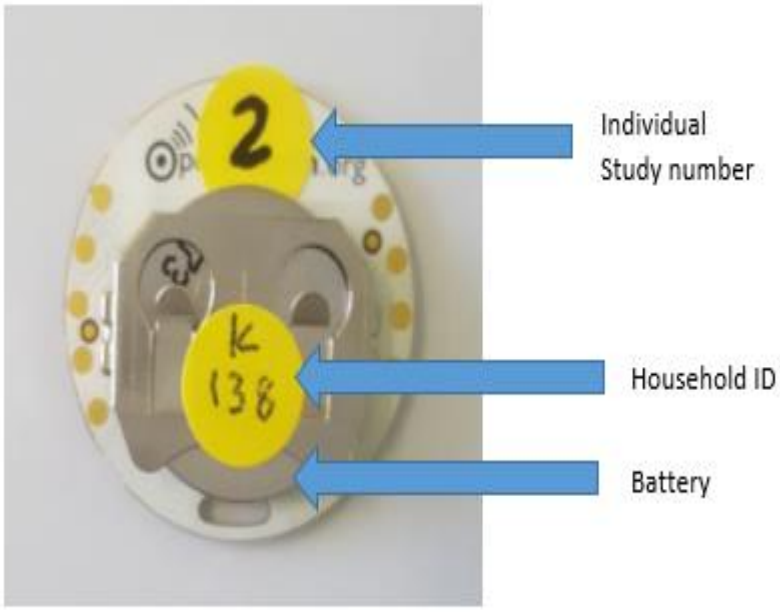


Figure B.3*: Rear view of tag with participant study ID.*

For the participant tag shown in figure 6, the study ID is: K138-002. This type of double-labelling on the tag and the pouch ensure that the right individual receives the right tag. The participants know their study ID's and are encourage to only wear the pouches with their study ID's and not swap them around.

2. Noise

The proximity sensors are continuously exchanging power packets with sensors that are in close range. If the tags are not worn, but are in close proximity to each other, they will continue to log proximity events which is not a true reflection of the situation. This is common in the evenings when the

participants are sleeping and may have taken the tags off and placed them together. This will generate noisy readings which can be filtered out using the time-stamp and accelerometer readings. A second control measure that can be put in place, is storing the tags safely in foil pouches which will contain the RF-signals and help minimize noise. Figure B.4 shows an example of the foil pouch.

Figure B.4*: Front and rear view of the foil pouches*

The foil pouches are made of paper and wrapped with foil on the exterior. The paper inside is used to prevent the reflection of the RF-signals which could heat up the sensors. The impact of the foil pouch will be analyzed from the data.

3. Meta Data

Finally, a spread sheet containing all the meta data is compiled. This includes the tag ID on the proximity tag itself, linked to the participant study ID, their name and surname, age and gender. The name and surname will be omitted in the analysis for confidentiality purposes. Figure 6 below shows a snapshot of the meta data file.

A	B	C	D	E	F
Tag ID	Individual ID	Name	Surname	Age	Gender
664459	A112-001	Tsakani	Mbambo	44	F
798358	A112-003	Lucracia	Mambane	22	F
798354	A112-004	Surrender	Mbambo	7	M

Figure B. 5*: Meta data file for linking proximity tag ID's to the participants*

Appendix C – Raw Data Extraction

C1. Raw Data Extraction

In this section, the detailed process of raw data extraction is explained. The section begins with the system infrastructure, followed by data extraction and the preparation of the raw data for the next stage which is the data pre-processing.

C2. SocioPatterns System Infrastructure

This process relies exclusively on the proprietary software developed by the ISI Foundation in Italy and the library and specific processing details cannot be shared for protection of intellectual property. The system is developed on a Linux platform and the system infrastructure is shown in Figure C.1

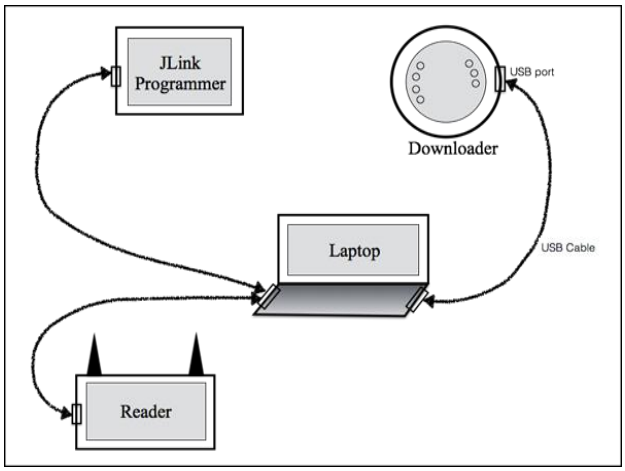


Figure C.1: ***System infrastructure of the SocioPatterns platform for the proximity study***

The system consists of a JLink programmer, a reader and downloader, all attached via USB connections to the laptop. The reader is only used to visualise the activity of the proximity tags in the real live environment during deployment, which will not be used in the study. The proximity tag is attached to the downloader which acts as a dongle. The programmer is a small pc board, with a microprocessor that is used to load a proprietary firmware (developed by the ISI Foundation in Italy) onto the proximity sensor. Linux commands are used to interact with the programmer and

downloader. The data is downloaded from the proximity tags using this infrastructure.

C3. Data Extraction

Data extraction relates to the download of data from the proximity sensors to produce the final raw csv file. Figure C.2 shows the flow diagram of this data extraction process.

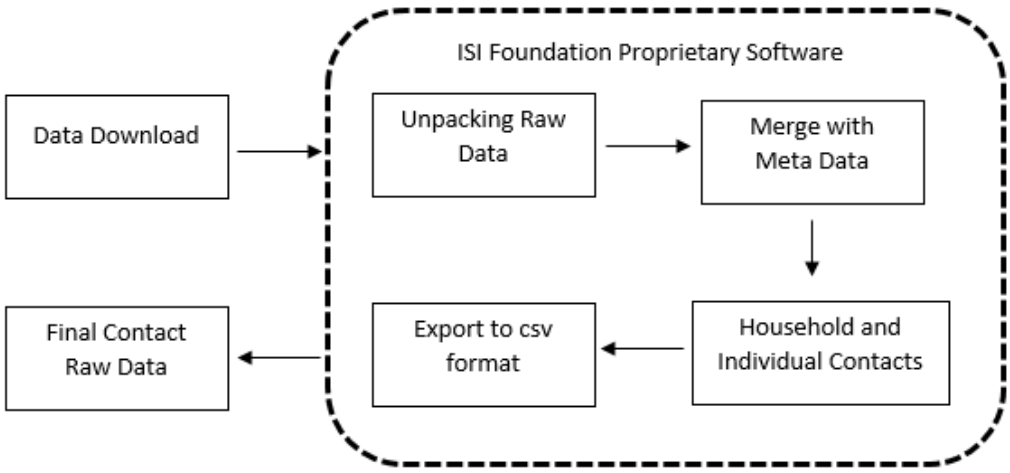


Figure C.2: ***Flow Diagram of Data Extraction Process***

The dotted part of the diagram shows the processing points in the data extraction process that are run using the proprietary libraries developed by SocioPatterns from the ISI Foundation in Italy. The process begins with the data download stage and ends with a raw csv file that is exported to be imported into Matlab.

Data Download: Every proximity tag used in the study is inserted onto a downloader shown in figure 4 and Linux commands are run on a terminal to begin the download process triggered by a python script developed by the ISI Foundation.

Specific details of this download process are outlined in Appendix C. Once the data has been downloaded from all the tags, the files are encrypted and ready for transfer to the ISI. This is done by sharing the files on Dropbox and this is the basis for encryption, to protect the data in transit through the network. Once the files have

been received, the files are decrypted into an hdf5 format and this file is imported into another script where the unpacking of the raw data begins.

Unpacking Raw Data: The hdf5 file is loaded into a Python Notebook where the unpacking of this data is done. The hdf5 file contacts the contact dataset and status dataset, which are extracted separately into a Dataframe.

	t	attenuation	Site Name_x	Household ID_x	Individual ID_x	Site Name_y	Household ID_y	Individual ID_y
0	2017-10-09 08:00:20+02:00	-69.0	Klerksdorp	K156	K156-004	Klerksdorp	K156	K156-005
1	2017-10-09 08:01:20+02:00	-67.0	Klerksdorp	K156	K156-001	Klerksdorp	K156	K156-002
2	2017-10-09 08:18:20+02:00	-67.5	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
3	2017-10-09 08:18:40+02:00	-66.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
4	2017-10-09 08:19:00+02:00	-69.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001

Figure C.3: ***Contact Dataset Extracted from the hdf5 file***

	t	attenuation	Site Name_x	Household ID_x	Individual ID_x	Site Name_y	Household ID_y	Individual ID_y
0	2017-10-09 08:00:20+02:00	-69.0	Klerksdorp	K156	K156-004	Klerksdorp	K156	K156-005
1	2017-10-09 08:01:20+02:00	-67.0	Klerksdorp	K156	K156-001	Klerksdorp	K156	K156-002
2	2017-10-09 08:18:20+02:00	-67.5	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
3	2017-10-09 08:18:40+02:00	-66.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
4	2017-10-09 08:19:00+02:00	-69.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001

Figure C.3, shows the Dataframe in Python created from the hdf5 file. In this Dataframe, the contact data is found with 9 fields. The figure has been explained in chapter two of this proposal, but the important point to highlight here is how the merging of the metadata file is done. The *uid1* an *uid2* fields are the hardware IDs of the sensors, which are merged with *id1* and *id2* (sensor IDs), during the download stage. It is these IDs i.e. *id1* and *id2* that are merged with the metadata, explained in the next paragraph.

Meta Data:

The meta data is important for analysis and interpretation of the proximity sensor data. The meta data in the proximity pilot study includes the sensor IDs, the household ID of each individual, the individual study ID, the site name and the age of the individual. The important data is the Household IDs and the Individual IDs, because the raw data can be grouped according to these IDs.

	t	attenuation	Site Name_x	Household ID_x	Individual ID_x	Site Name_y	Household ID_y	Individual ID_y
0	2017-10-09 08:00:20+02:00	-69.0	Klerksdorp	K156	K156-004	Klerksdorp	K156	K156-005
1	2017-10-09 08:01:20+02:00	-67.0	Klerksdorp	K156	K156-001	Klerksdorp	K156	K156-002
2	2017-10-09 08:18:20+02:00	-67.5	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
3	2017-10-09 08:18:40+02:00	-66.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001
4	2017-10-09 08:19:00+02:00	-69.0	Klerksdorp	K104	K104-004	Klerksdorp	K104	K104-001

Figure C.4: ***Merged Contact Dataset***

Figure C.4, shows the merged contact dataset, which is very important towards understand how the household and individual contacts are separated. Following figure 6, the Site Name, Household ID and Individual IDs are important in the separation of the dataset into the various household and individual contacts explained in the next paragraph.

Household and Individual Contacts: From the merged contact dataset, there are two fields for both Household ID and Individual ID. These show the two individuals who are in contact. The dataset is further split according to the site name, household ID and then individual ID, before exported to csv. It is the individual contacts that are the most important at this stage, as the focus of this master's research is the cleaning of false contact events from the true contact events.

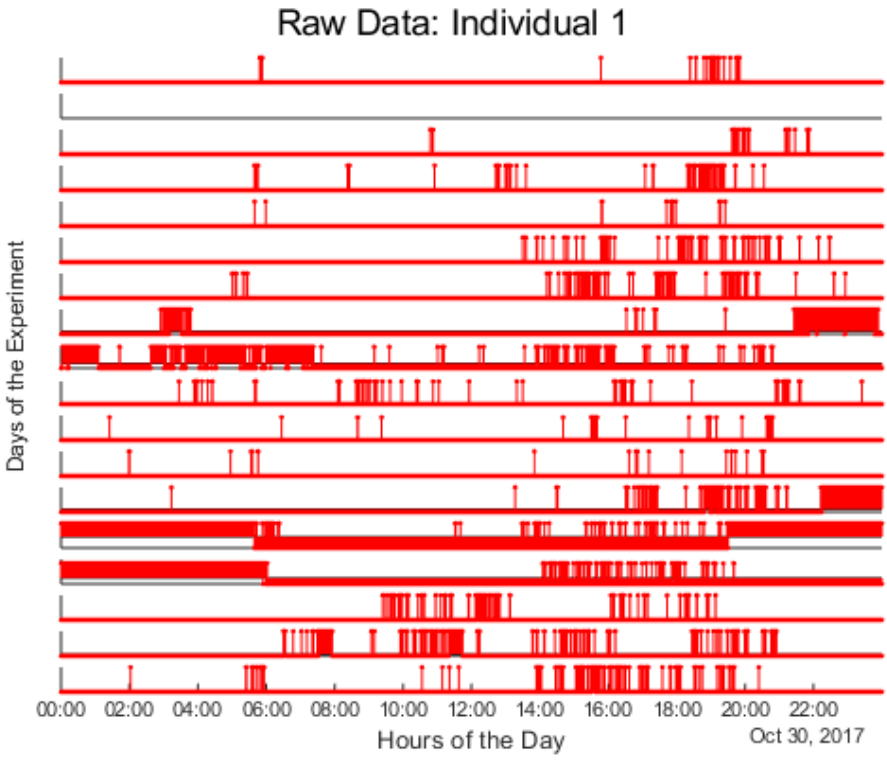
t	attenuation	Individual ID_x	Individual ID_y
2017-10-11 13:44:40+02:00	-65.0	A134-007	A134-001
2017-10-11 13:45:00+02:00	-63.0	A134-007	A134-001
2017-10-11 13:45:00+02:00	-50.0	A134-001	A134-003
2017-10-11 13:45:40+02:00	-58.0	A134-006	A134-001

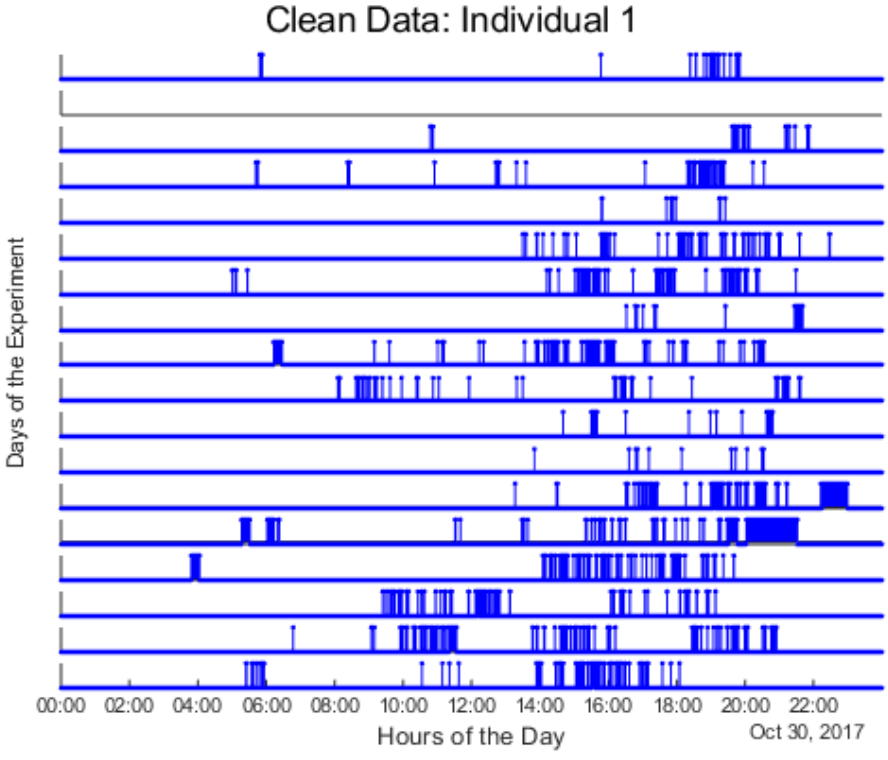
Figure C.5: ***Snapshot of the Final Raw Data File with Power Attenuation Values Between Two Tags at a Time t***

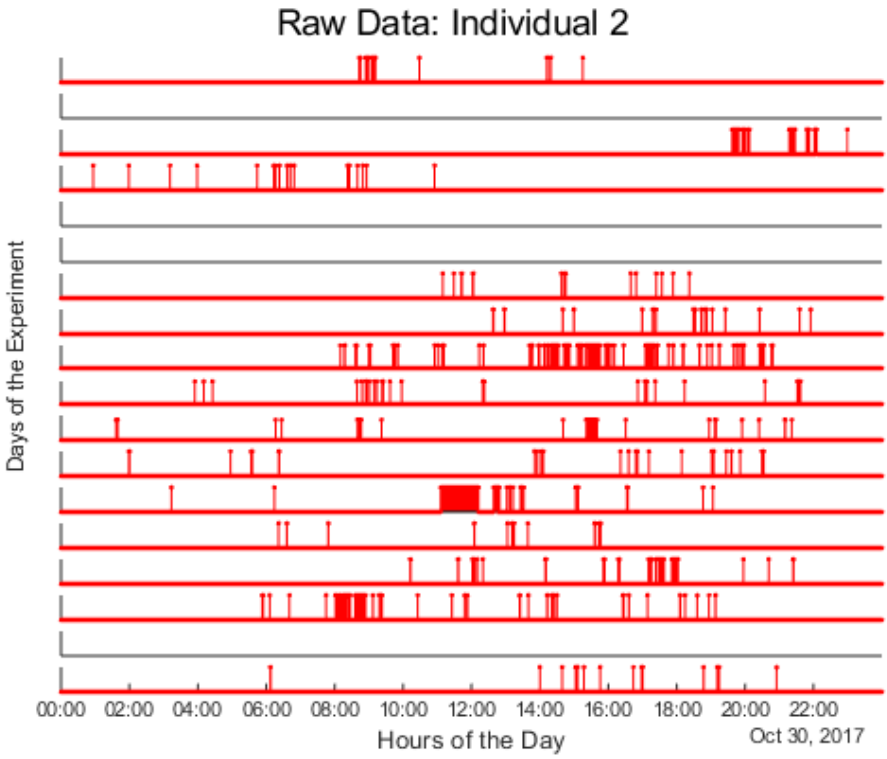
The data in Figure C.5, above shows the power attenuation values of contact which is the difference between the transmitted and received power, between any two pairs of tags, denoted by *Individual ID_x* and *Individual ID_y*. Each row in figure 10 represents a logged contact event.

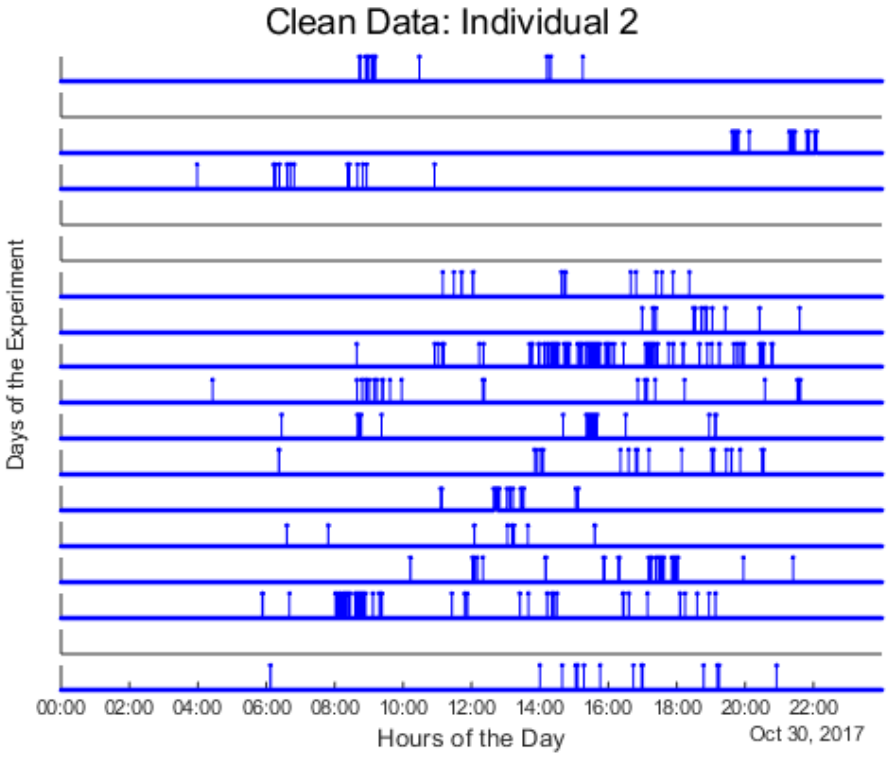
To re-iterate the challenges with data, the contact data set contains rows of contact events which are logged as power packets exchanged between two sensors at time t. The assumption is that individuals are wearing the sensors at all times, but this is not the case. The objective, which will be explored in the sections to follow, is to determine which of these contact events are true and which are not.

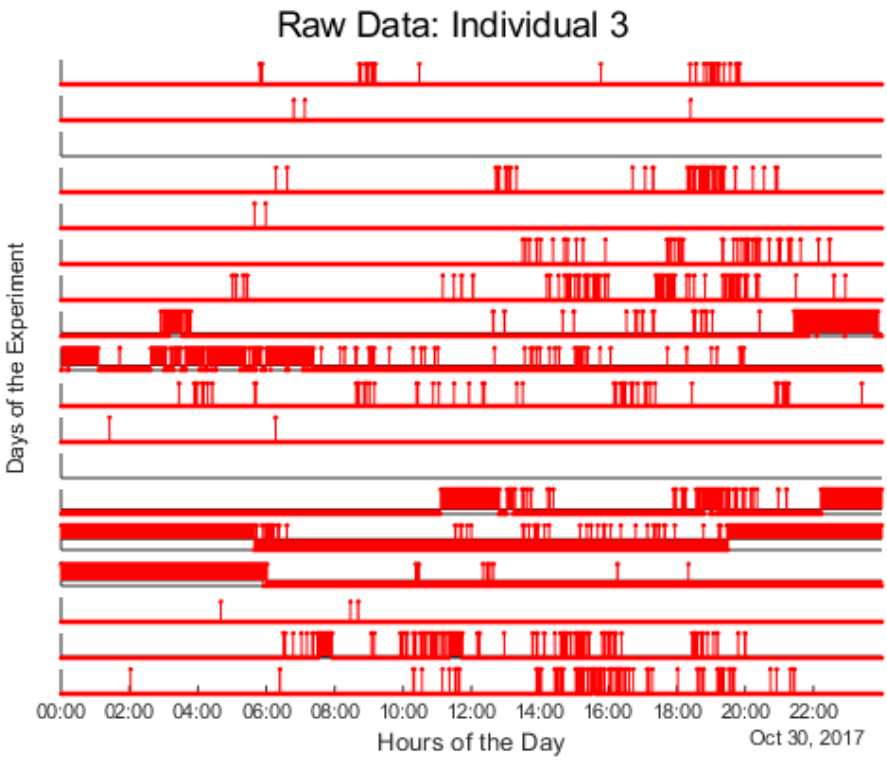
Appendix D – Raw and Cleaned Data

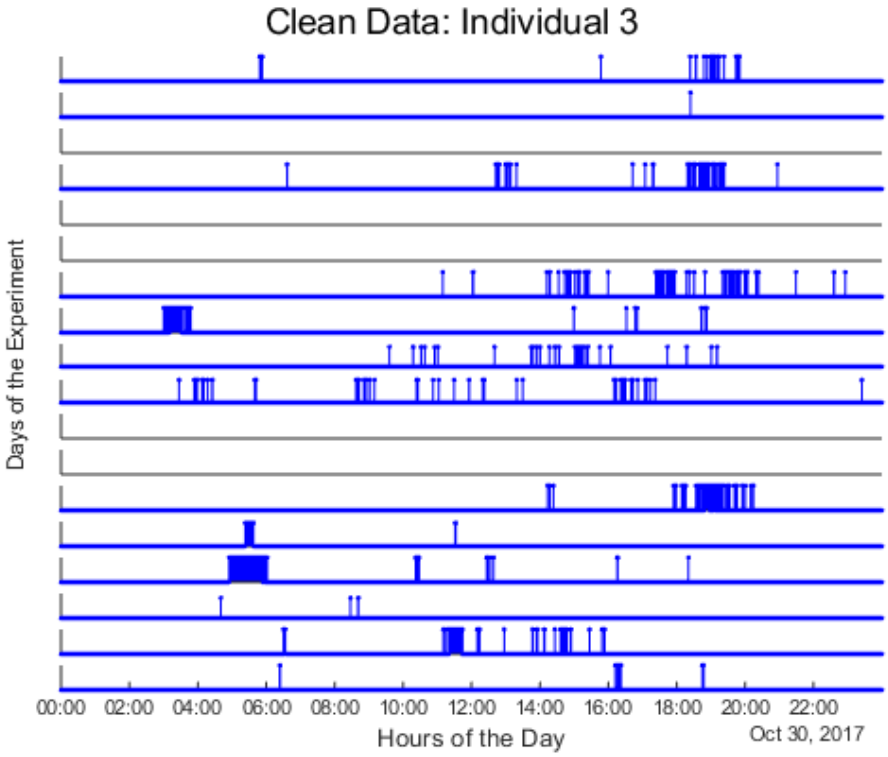


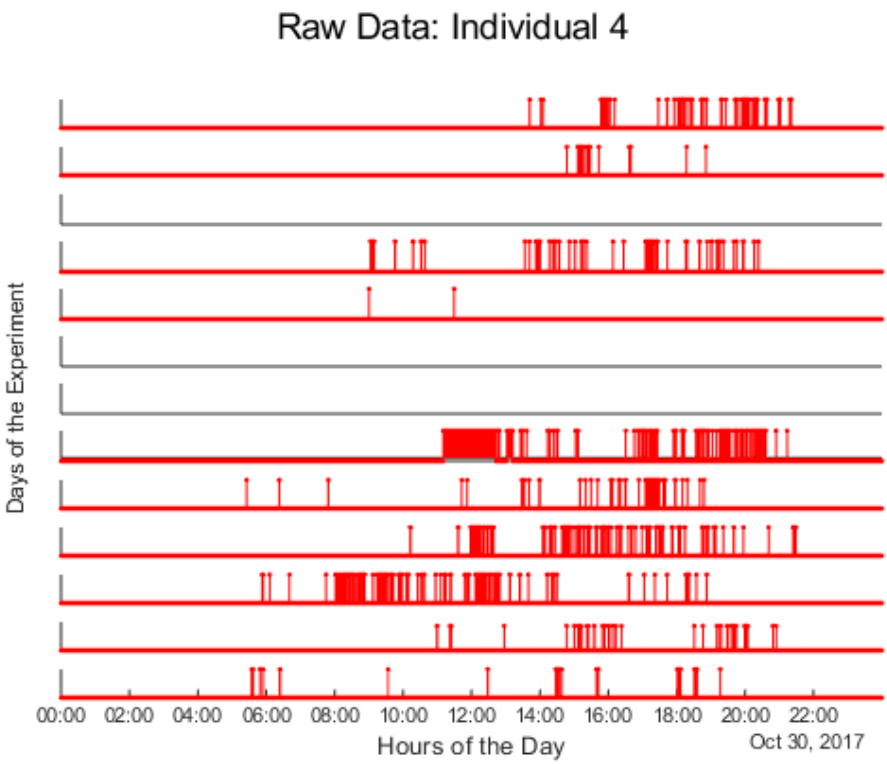


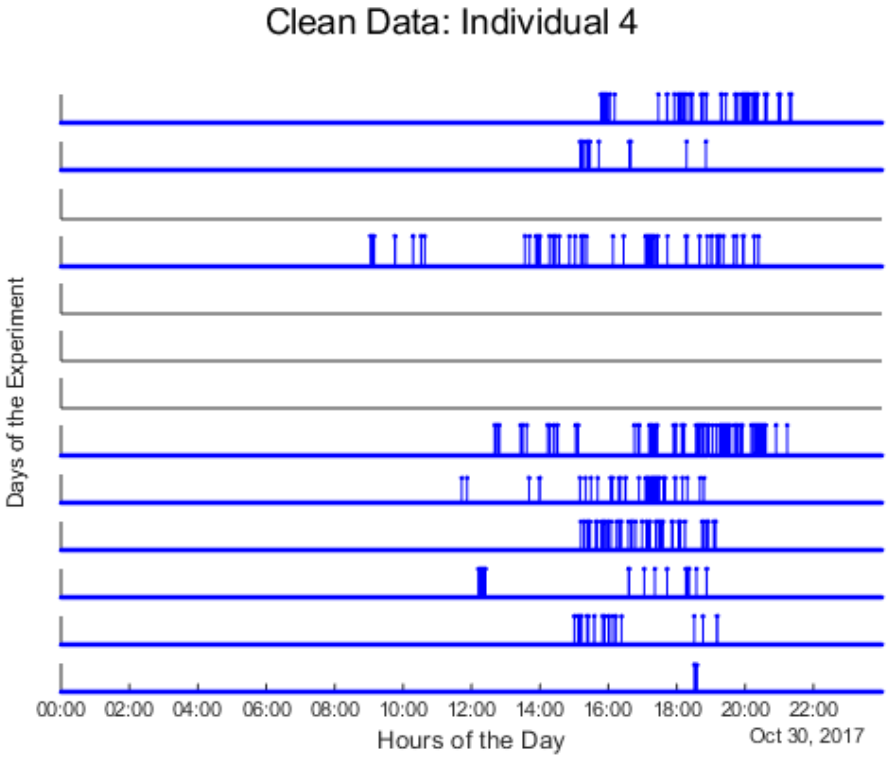

Raw Data: Individual 2
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

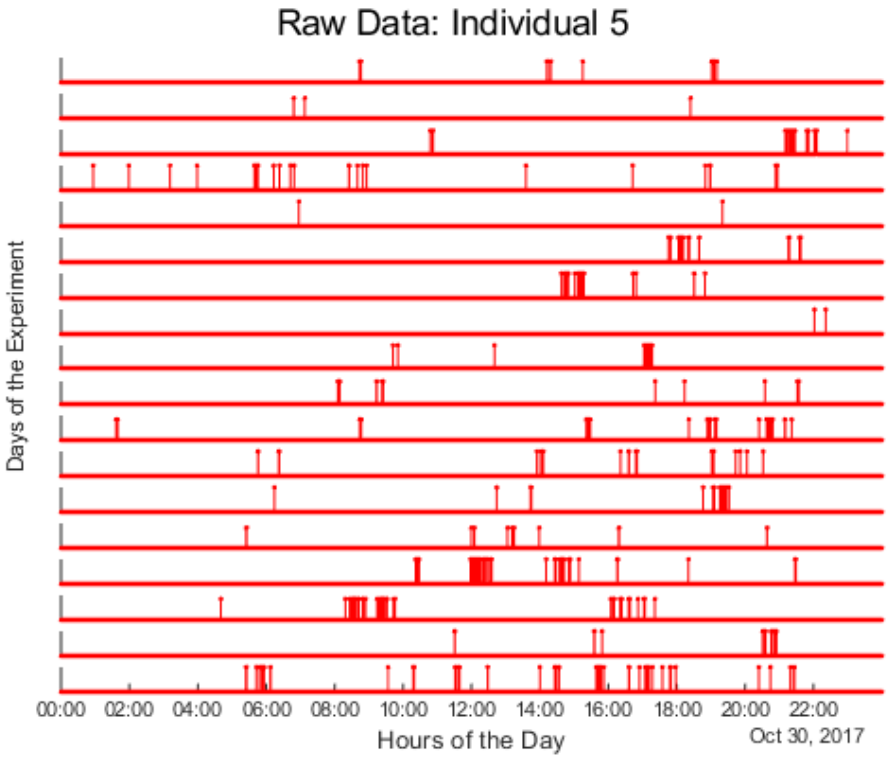

Clean Data: Individual 2
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

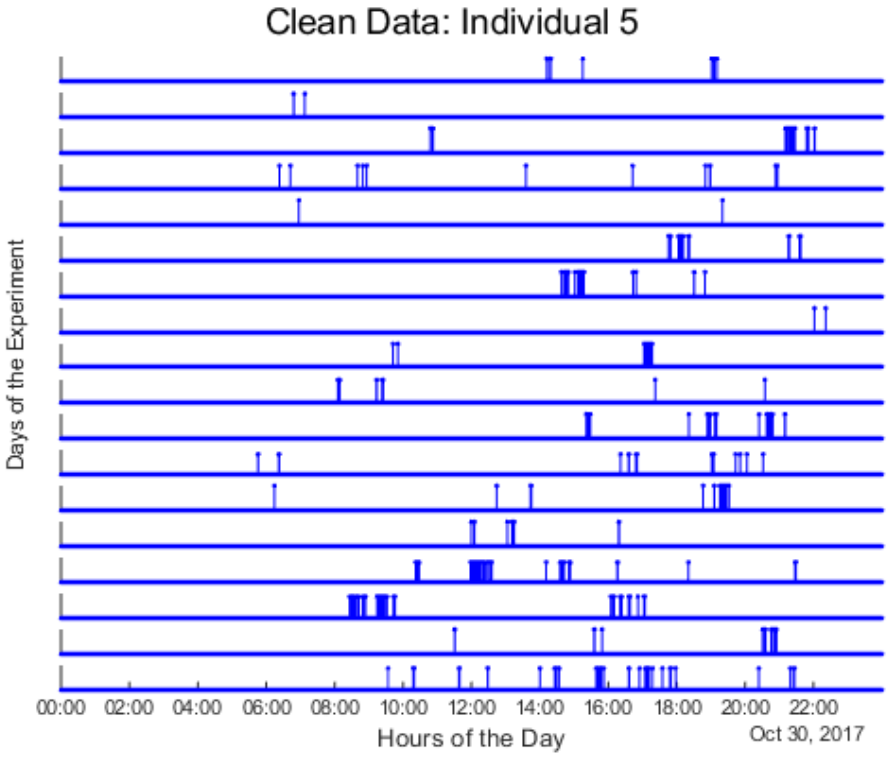

Raw Data: Individual 3
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

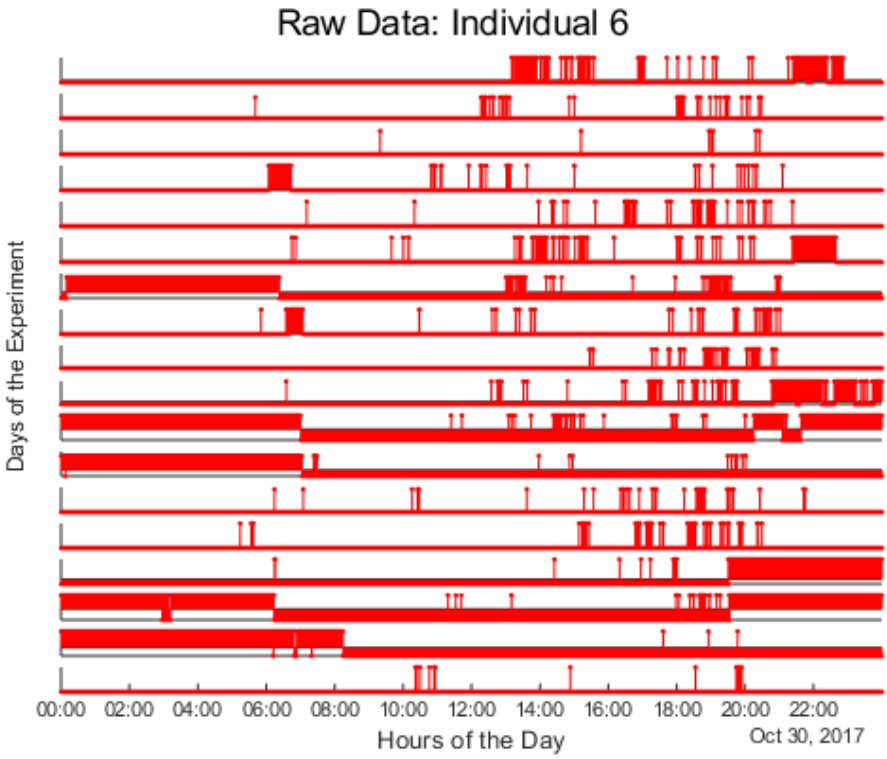

Clean Data: Individual 3
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

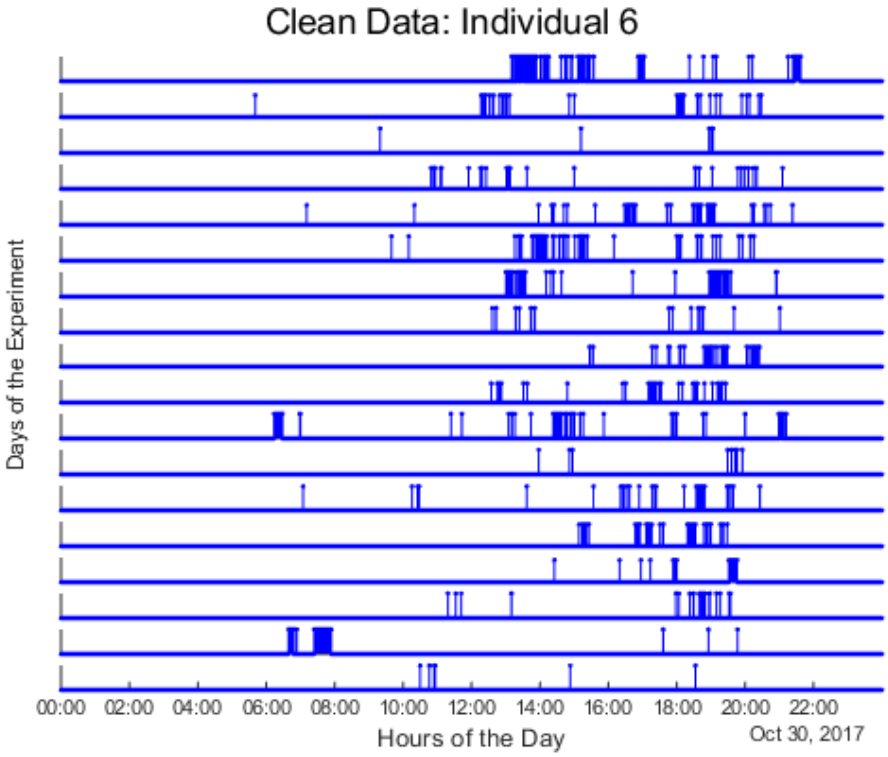

Raw Data: Individual 4
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

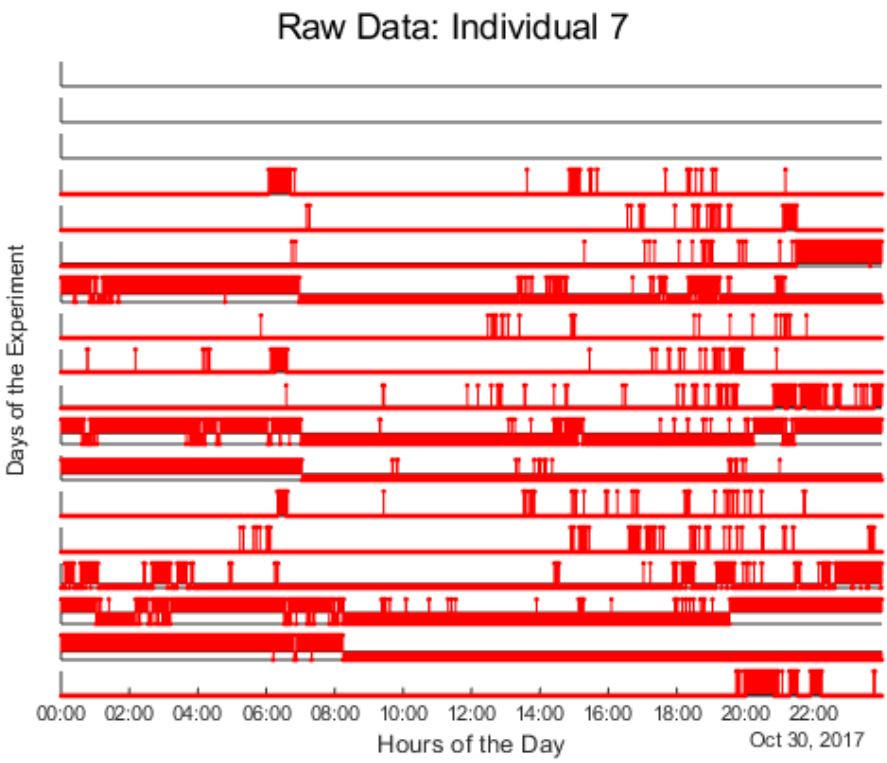

Clean Data: Individual 4
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

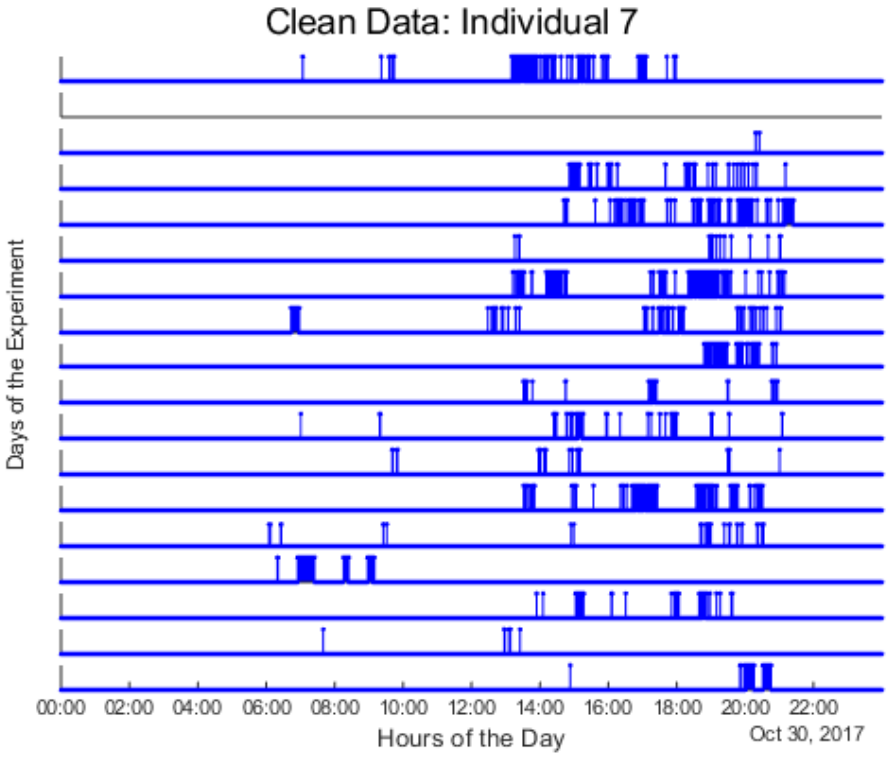

Raw Data: Individual 5
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

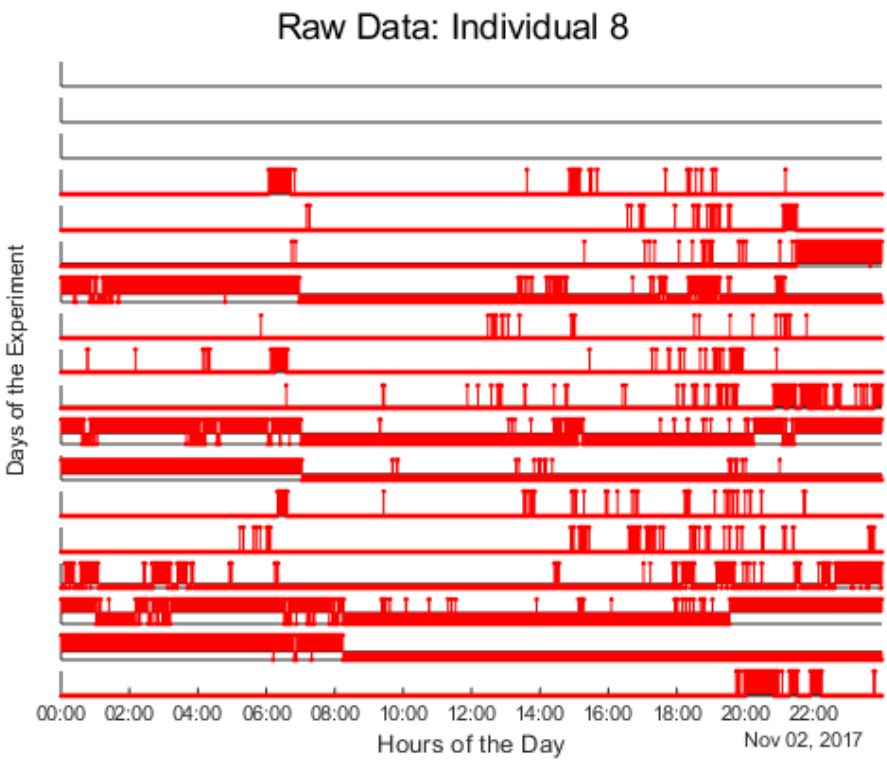

Clean Data: Individual 5
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

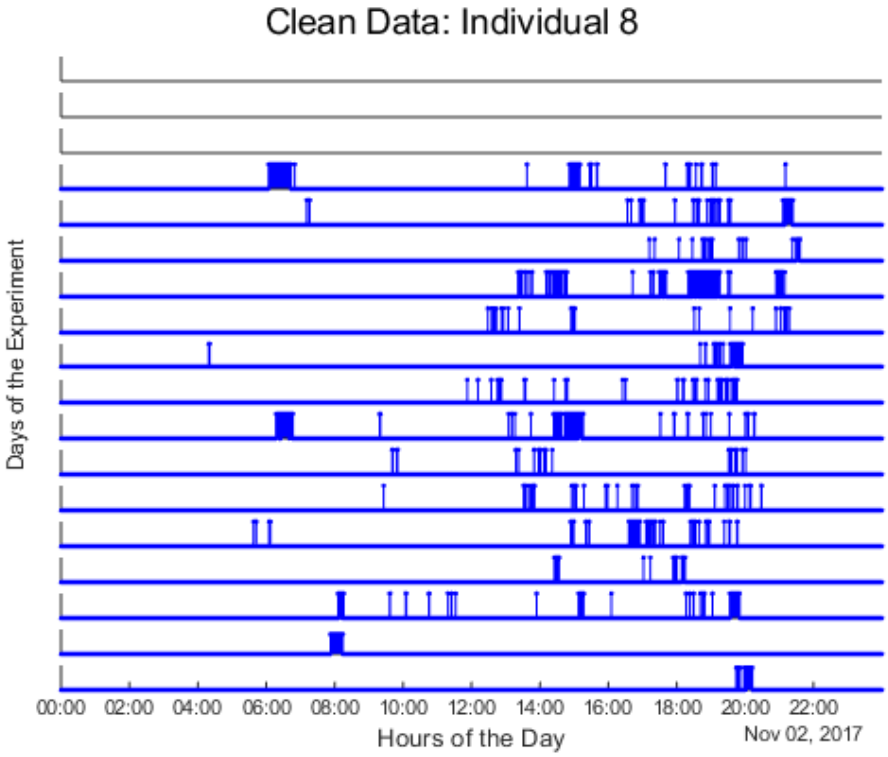

Raw Data: Individual 6
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

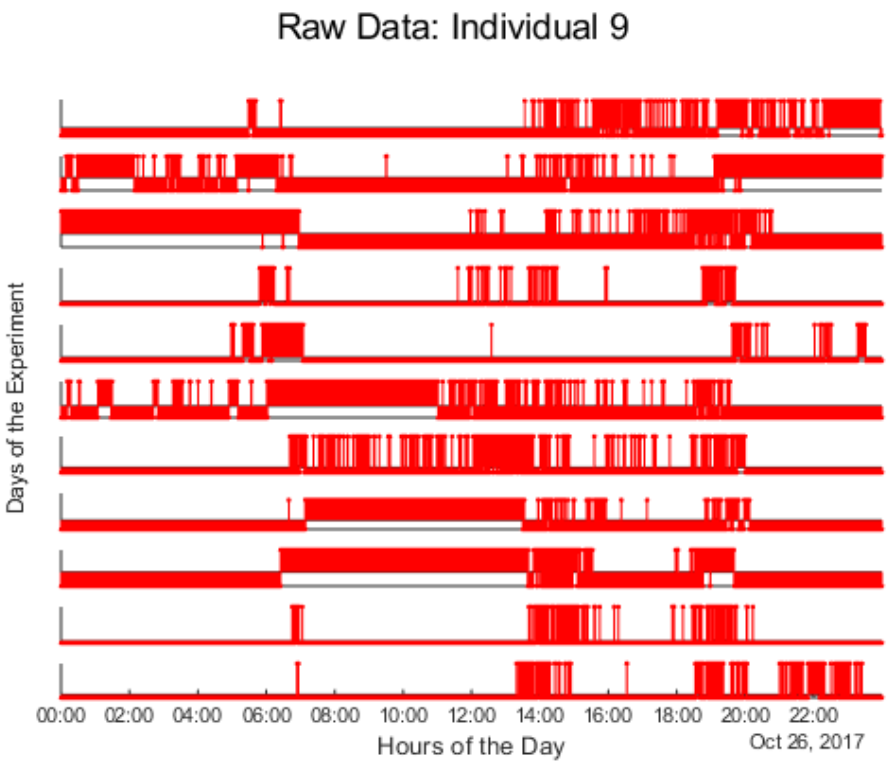

Clean Data: Individual 6
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

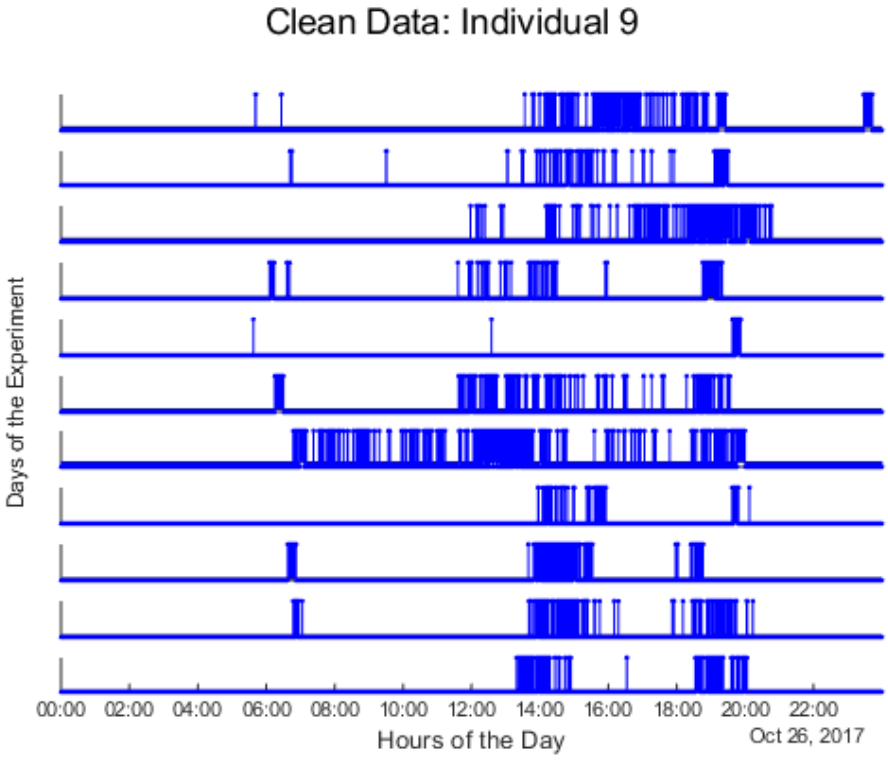

Raw Data: Individual 7
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

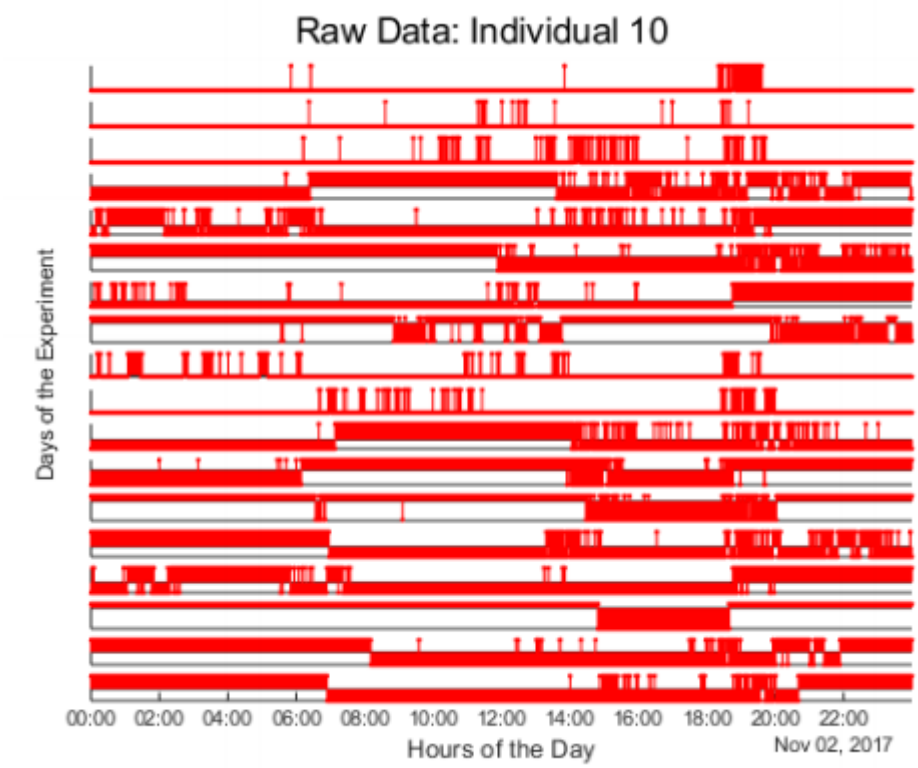

Clean Data: Individual 7
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

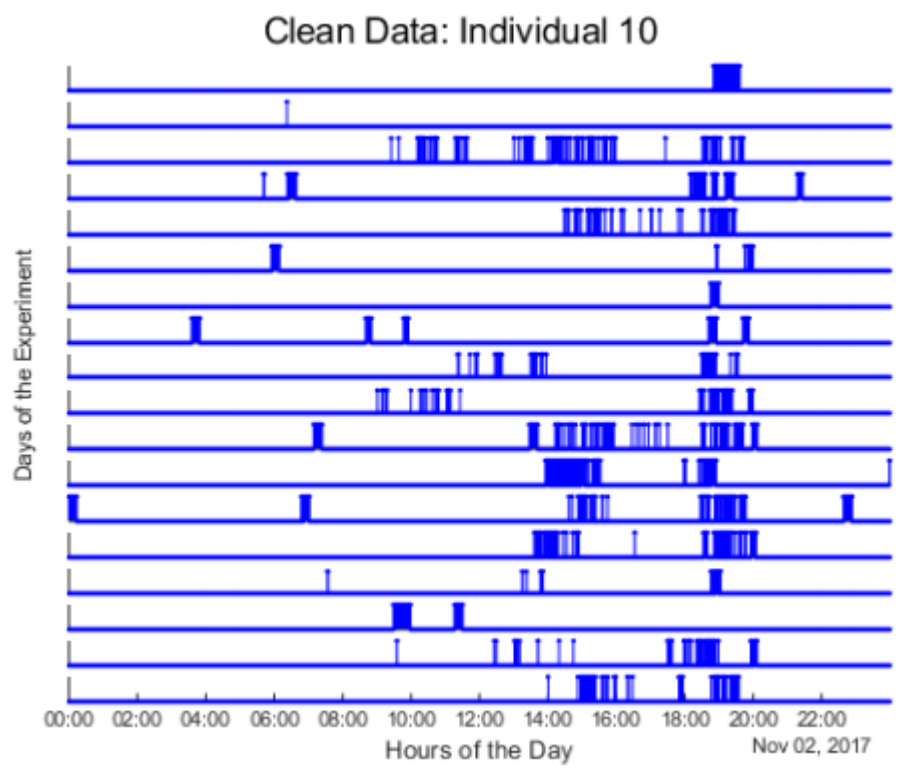

Raw Data: Individual 8
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Nov 02, 2017

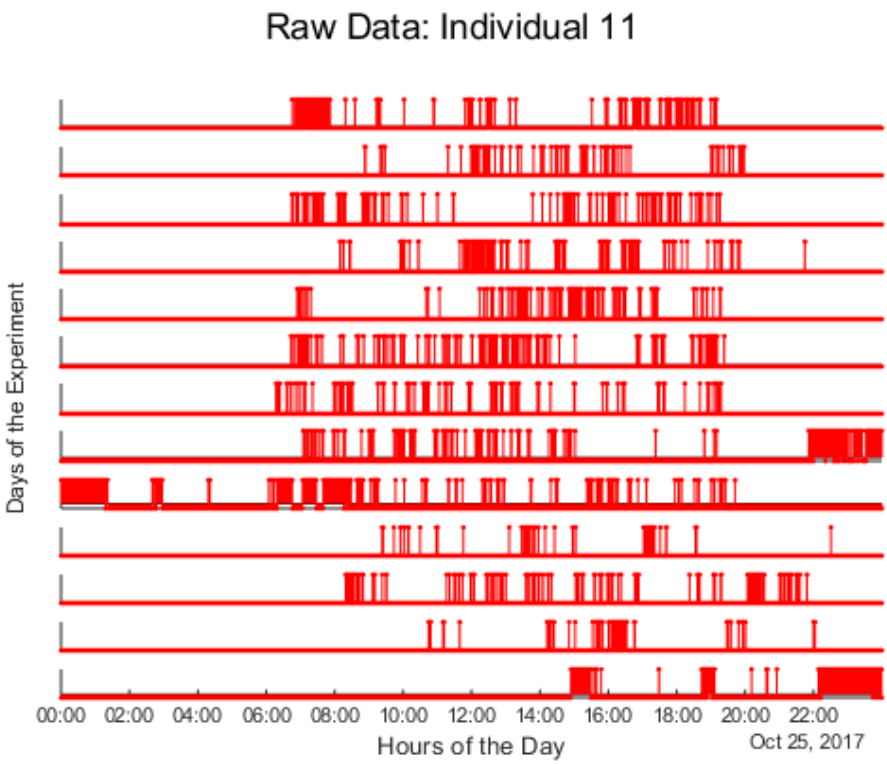

Clean Data: Individual 8
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Nov 02, 2017

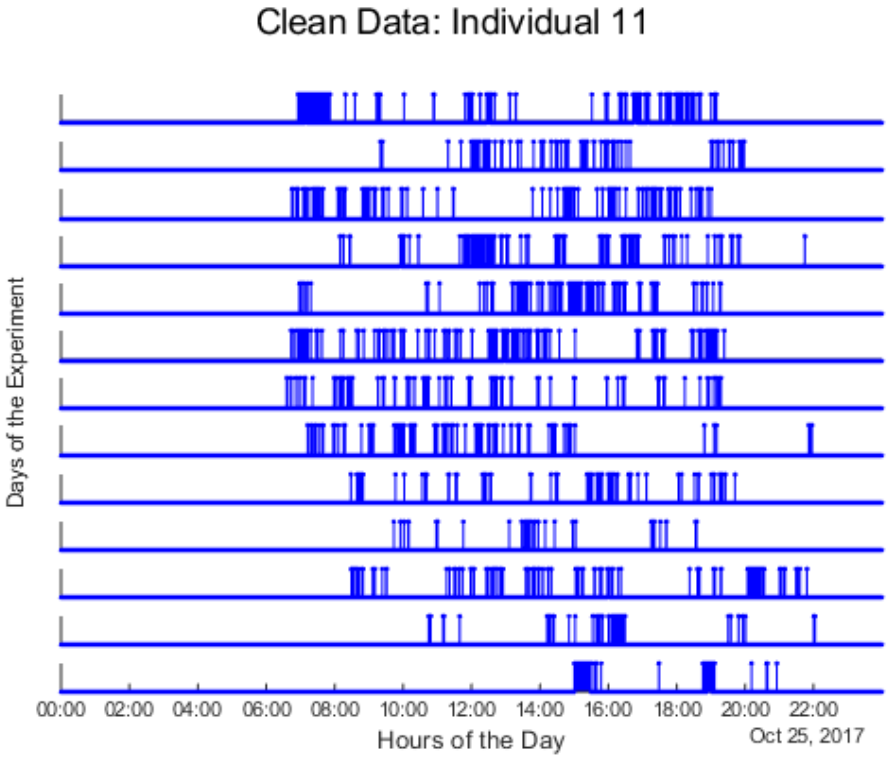

Raw Data: Individual 9
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 26, 2017

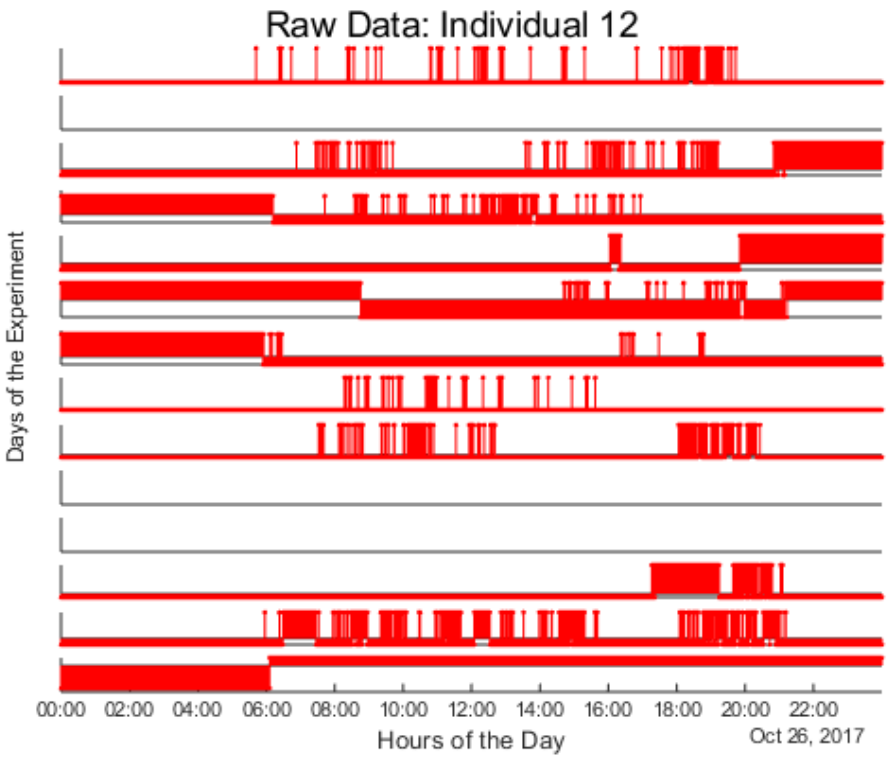

Clean Data: Individual 9
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 26, 2017

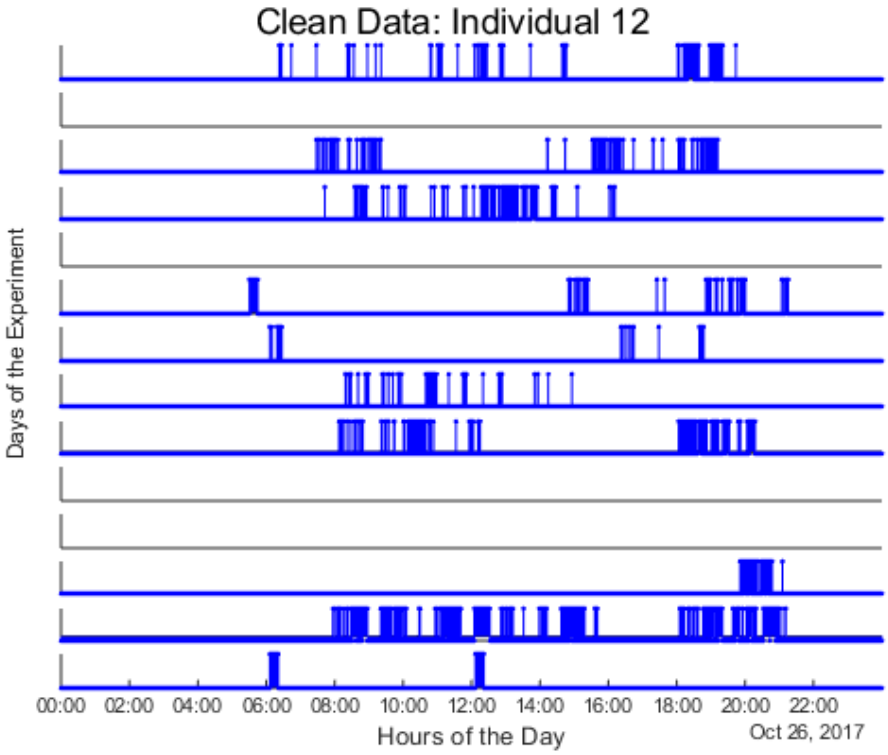

Raw Data: Individual 10
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Nov 02, 2017

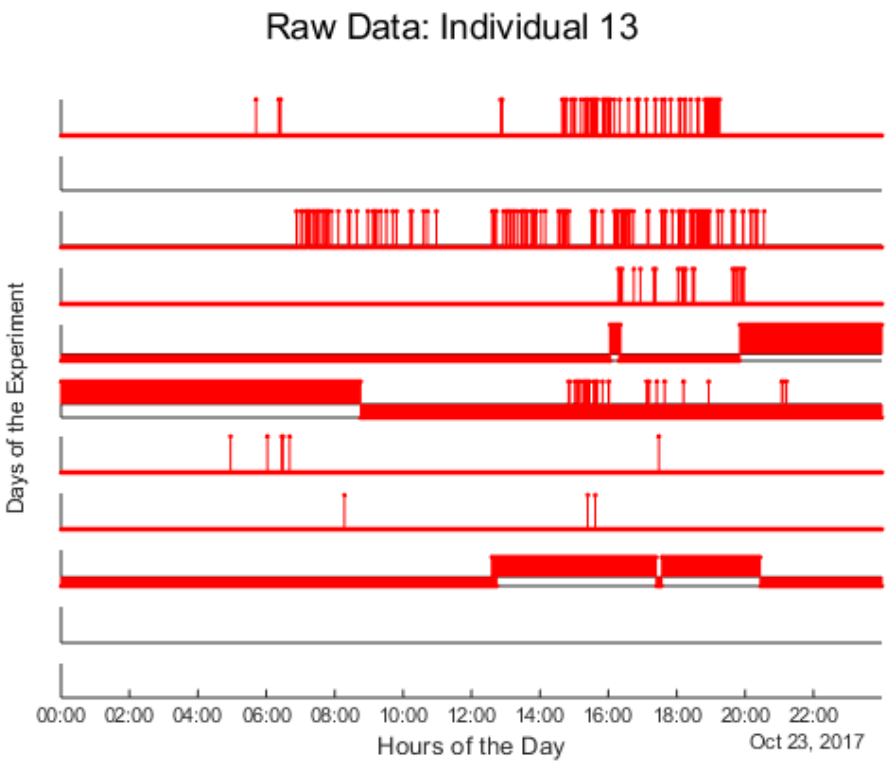

Clean Data: Individual 10
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Nov 02, 2017

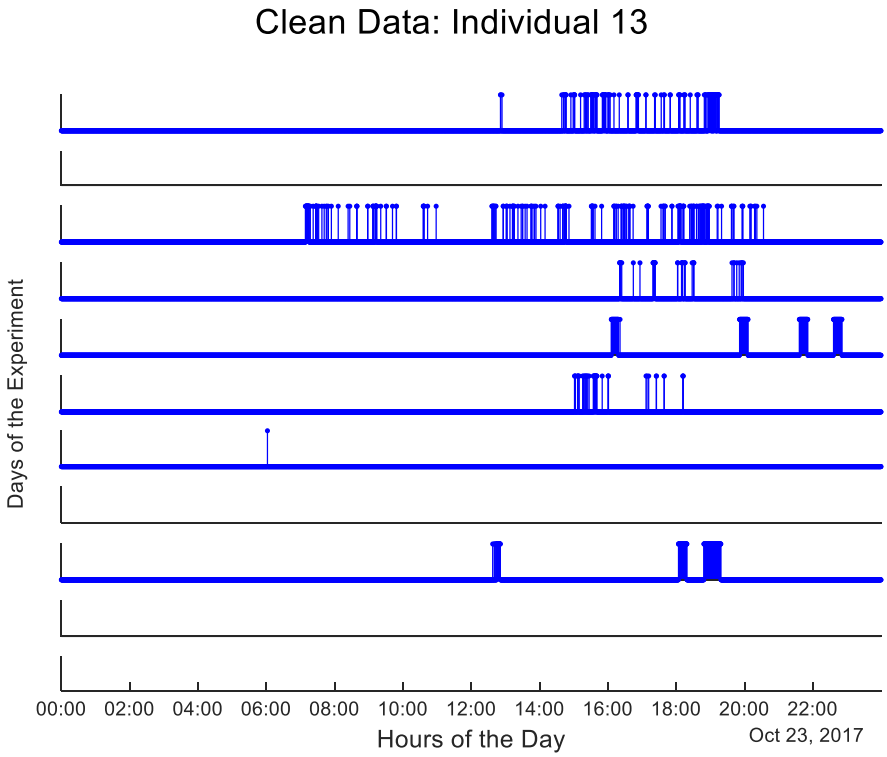

Raw Data: Individual 11
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 25, 2017

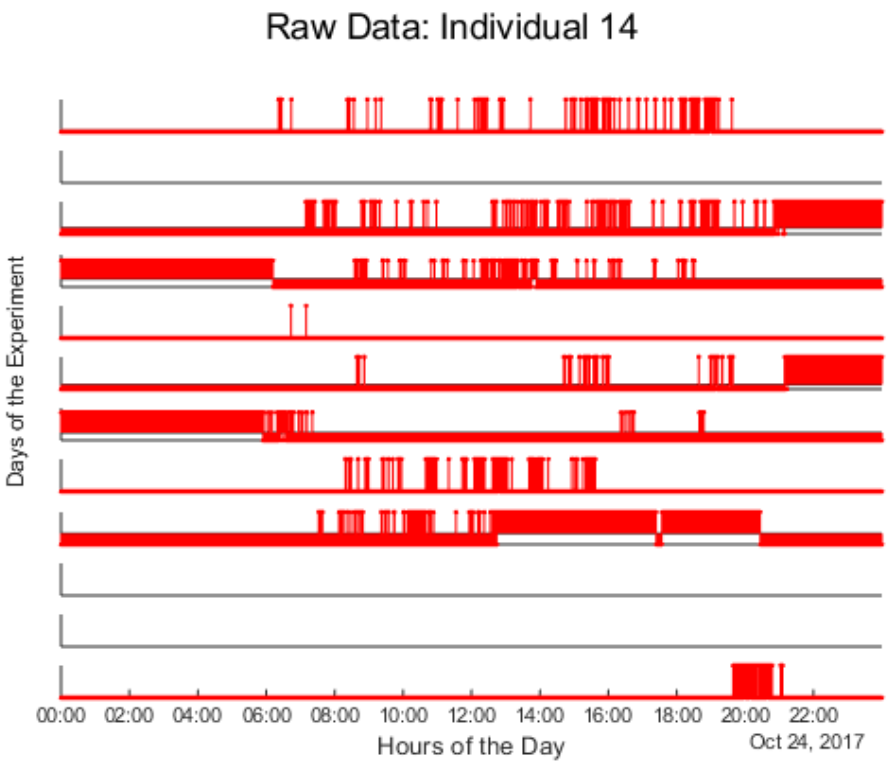

Clean Data: Individual 11
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 25, 2017

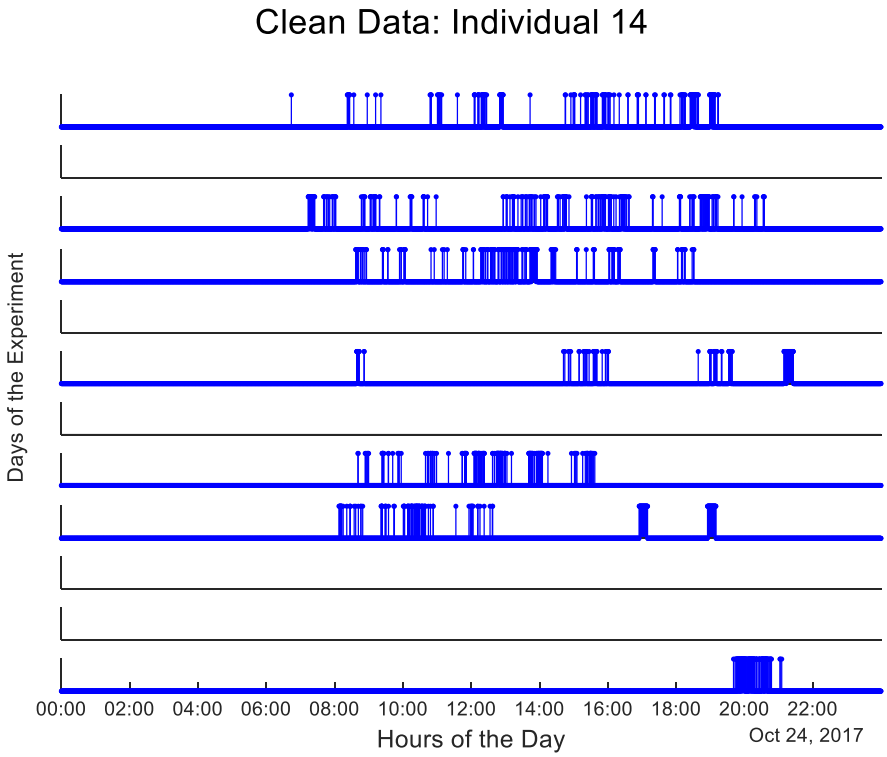

Raw Data: Individual 12
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 26, 2017

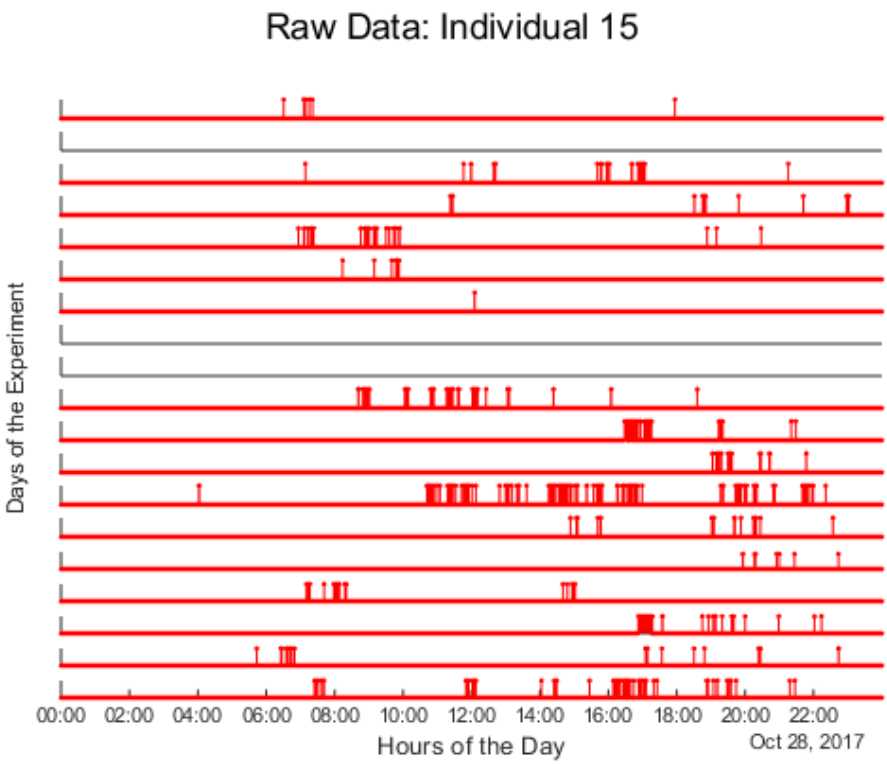

Clean Data: Individual 12
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 26, 2017

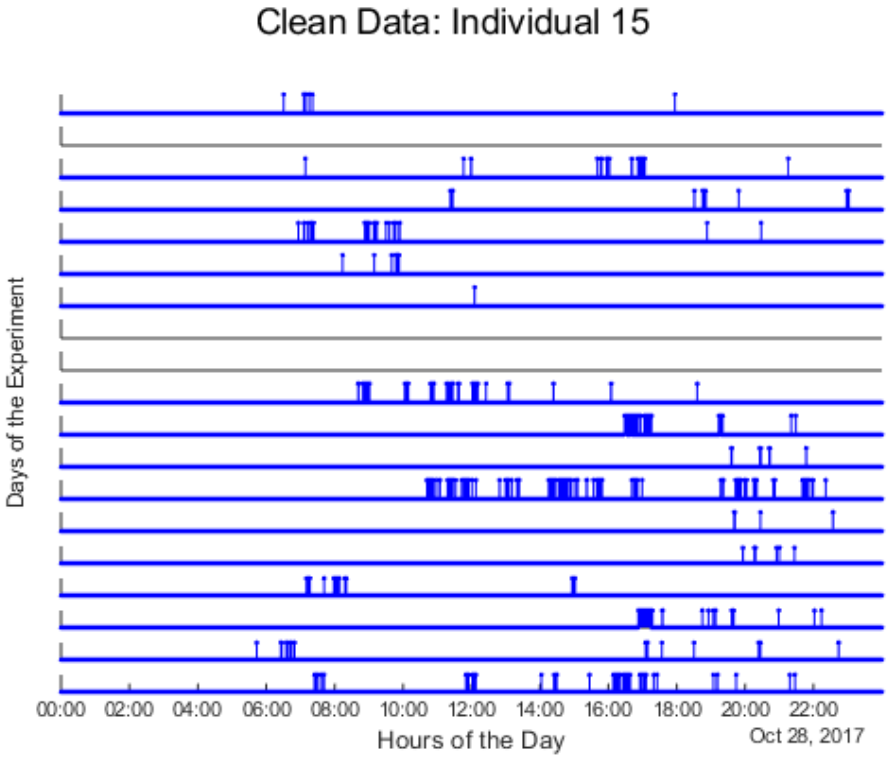

Raw Data: Individual 13
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 23, 2017

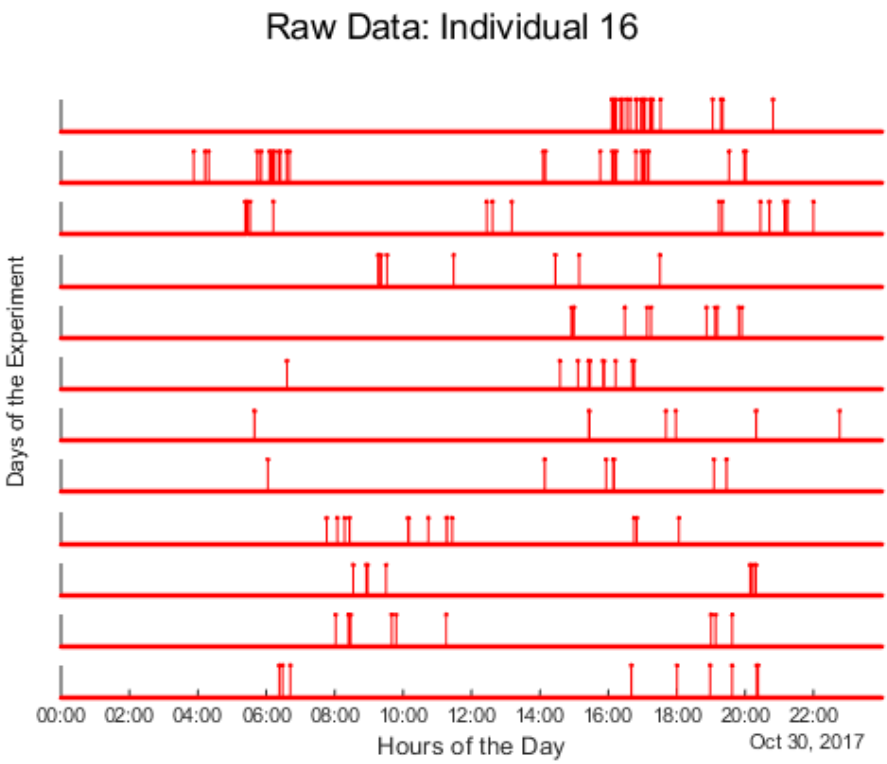

Clean Data: Individual 13
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 23, 2017

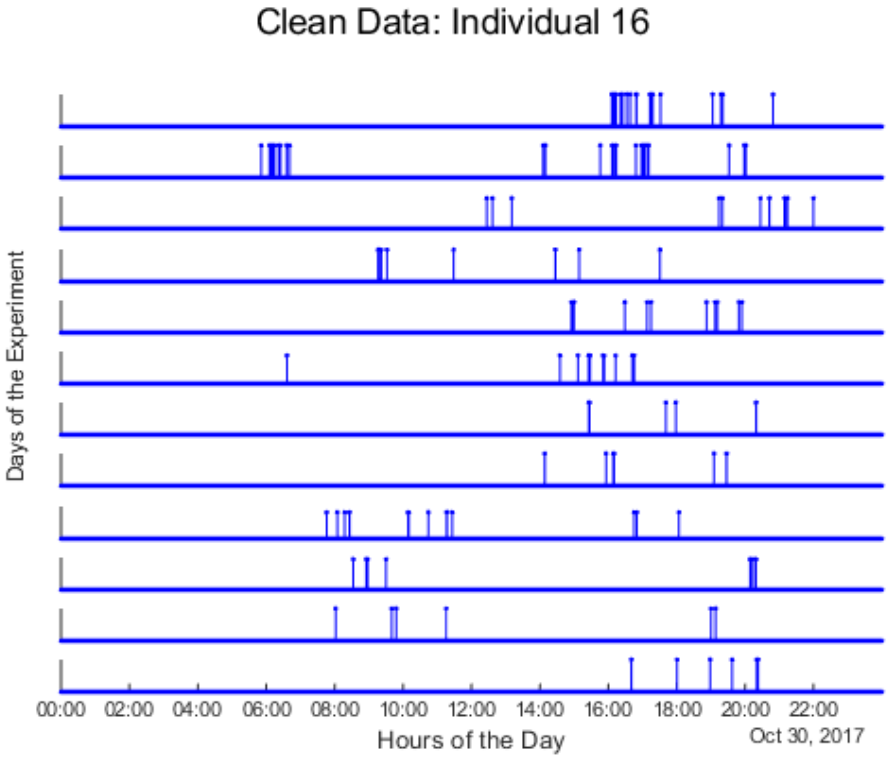

Raw Data: Individual 14
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 24, 2017

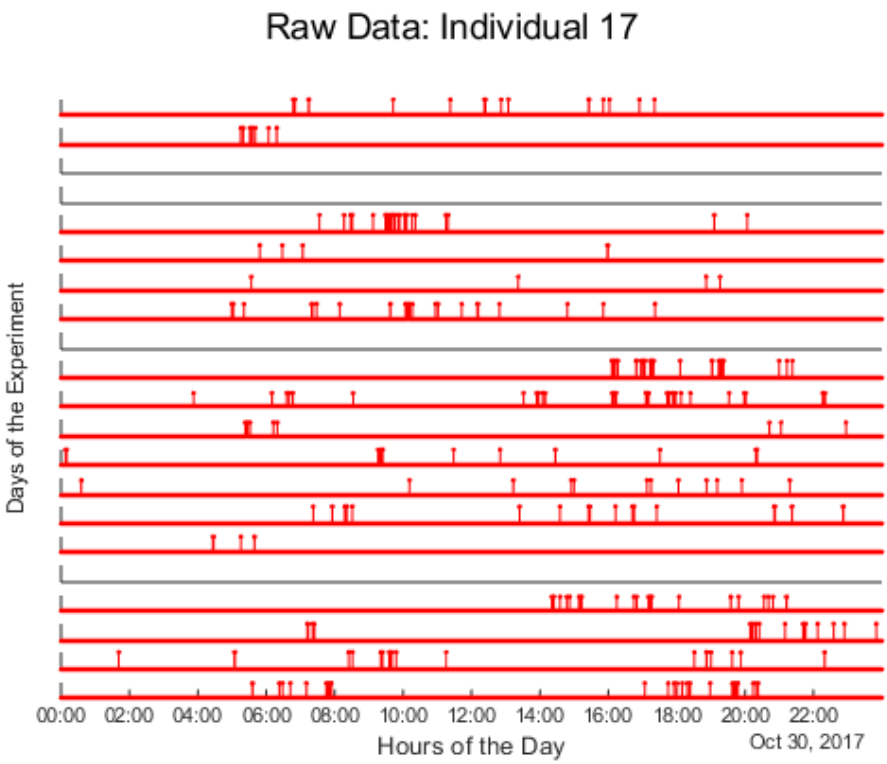

Clean Data: Individual 14
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 24, 2017

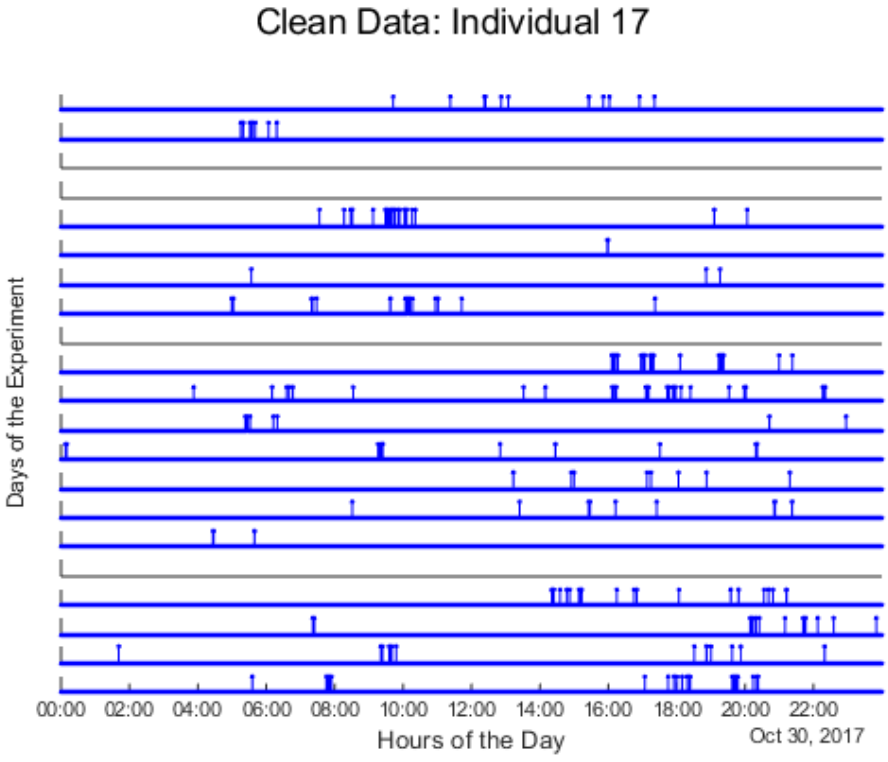

Raw Data: Individual 15
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 28, 2017

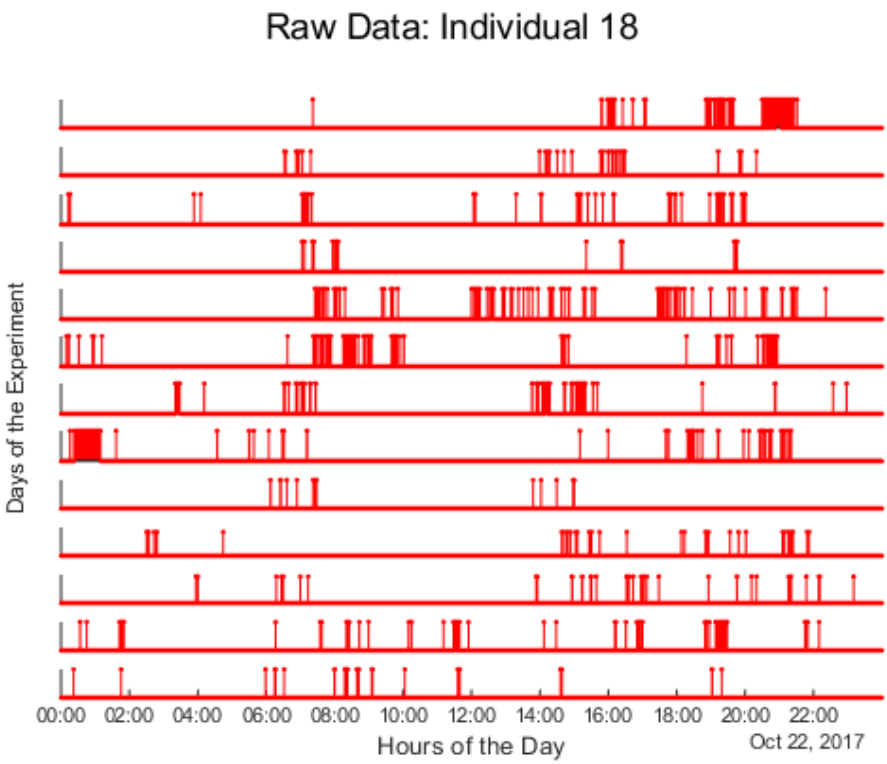

Clean Data: Individual 15
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 28, 2017

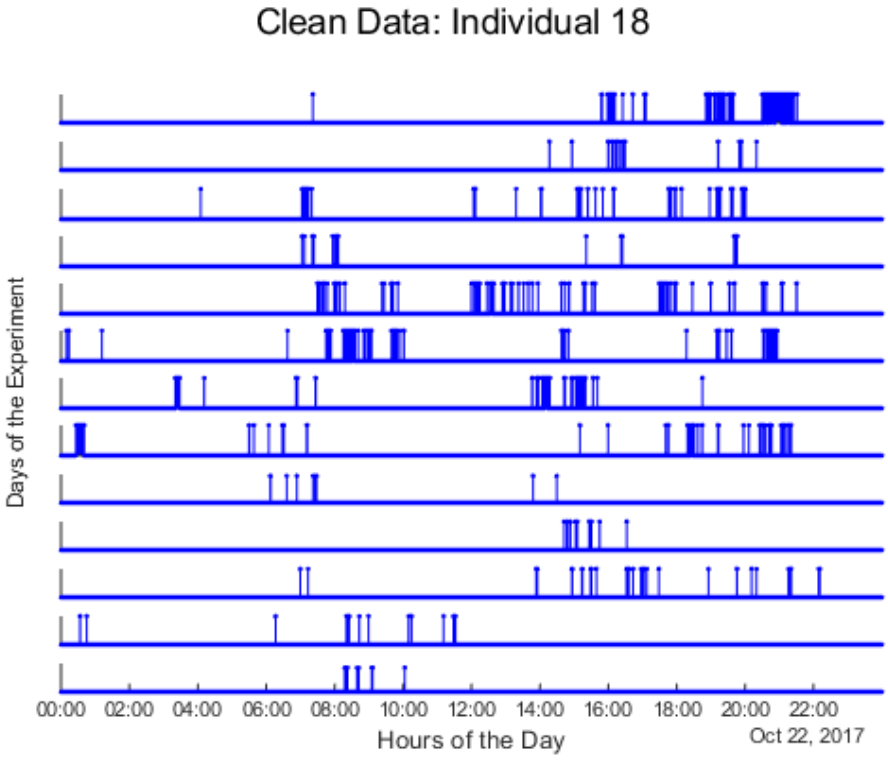

Raw Data: Individual 16
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

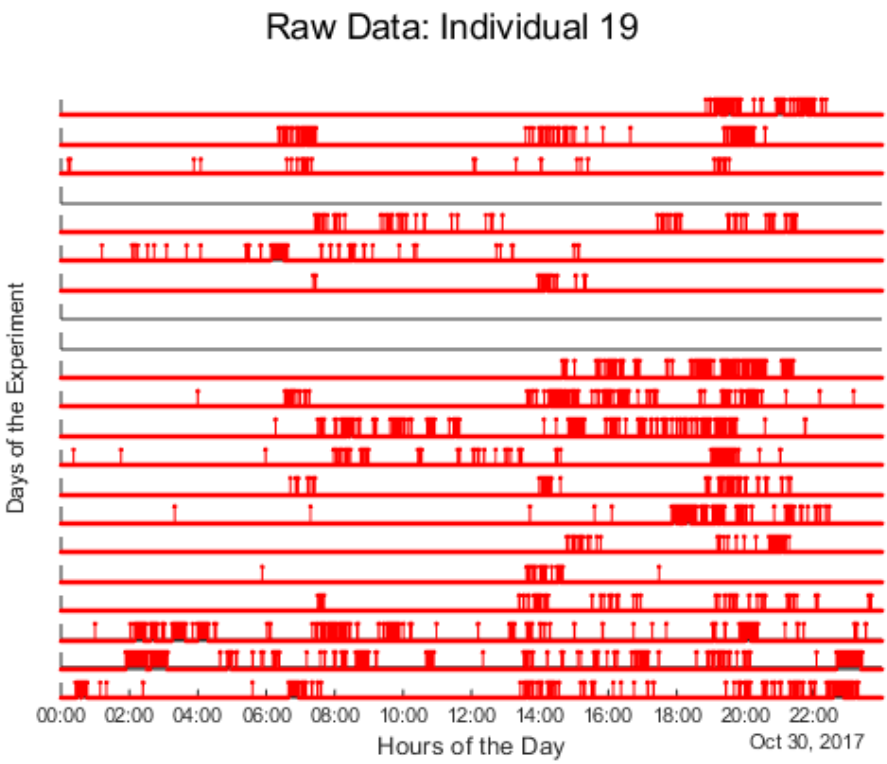

Clean Data: Individual 16
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

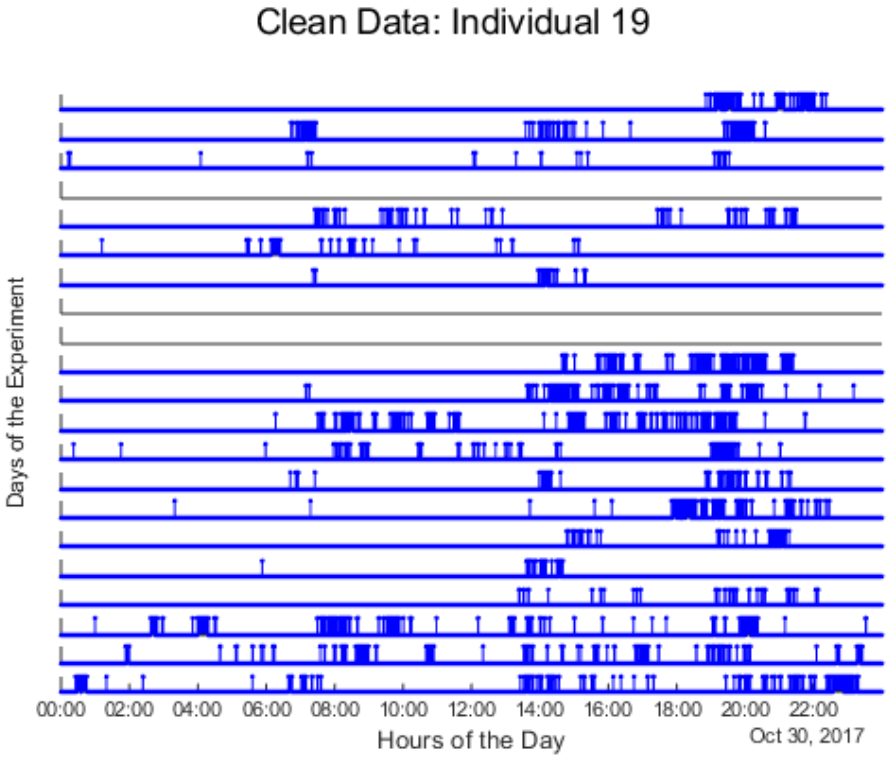

Raw Data: Individual 17
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

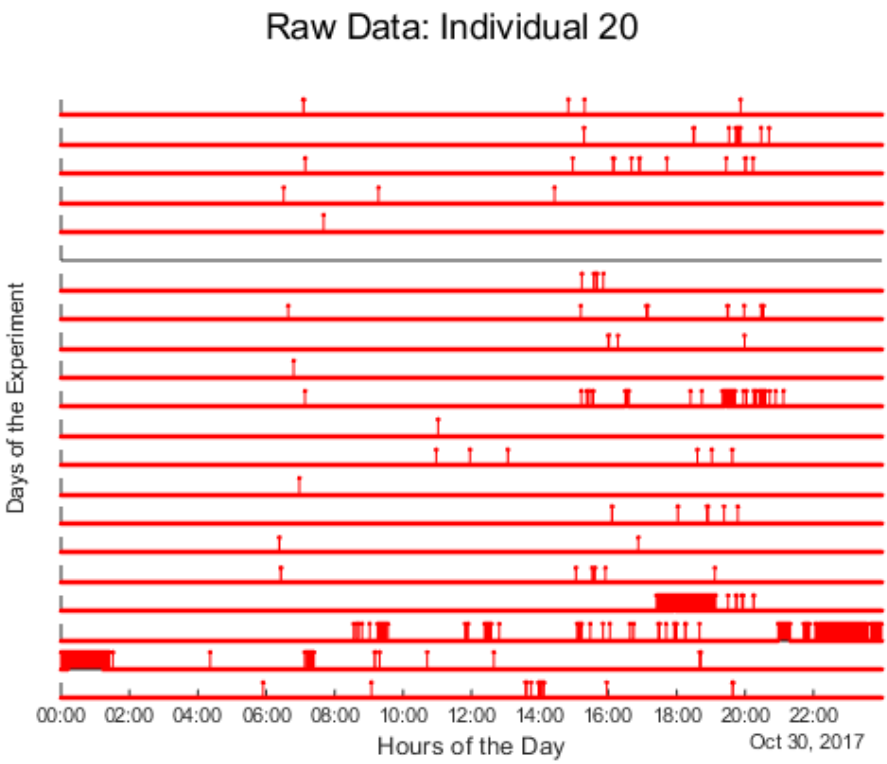

Clean Data: Individual 17
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

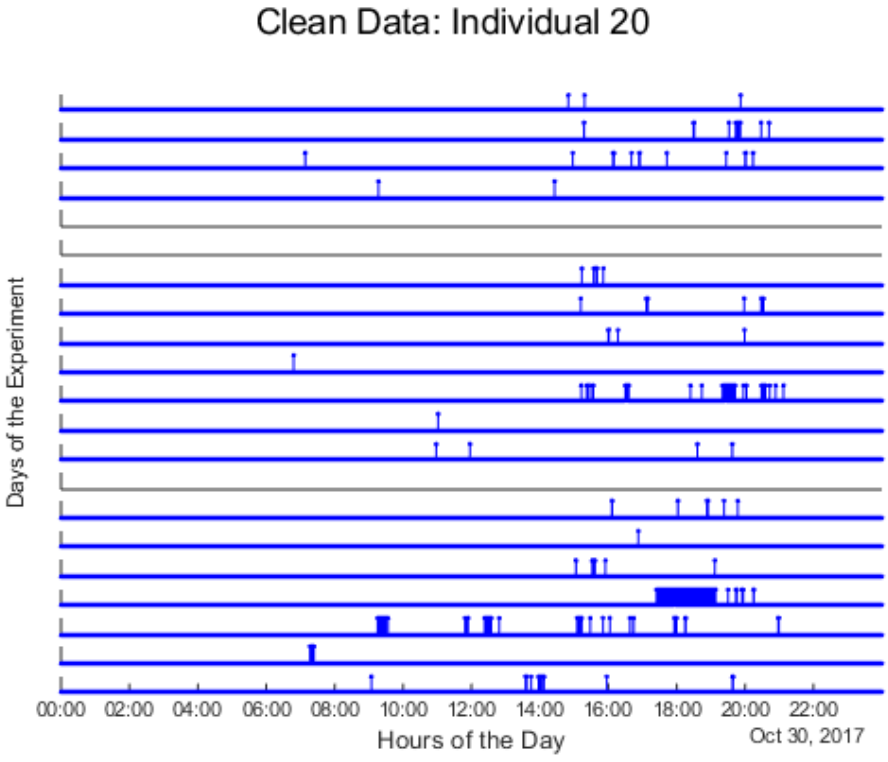

Raw Data: Individual 18
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 22, 2017


Clean Data: Individual 18
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 22, 2017


Raw Data: Individual 19
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017


Clean Data: Individual 19
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

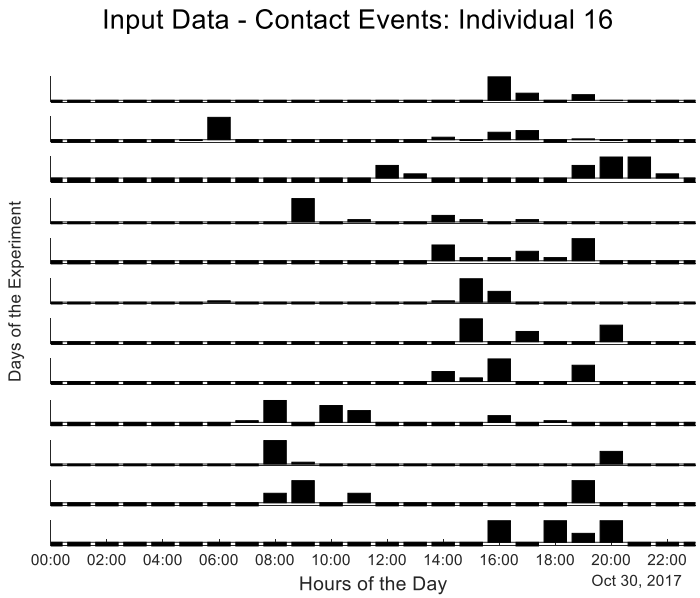

Raw Data: Individual 20
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

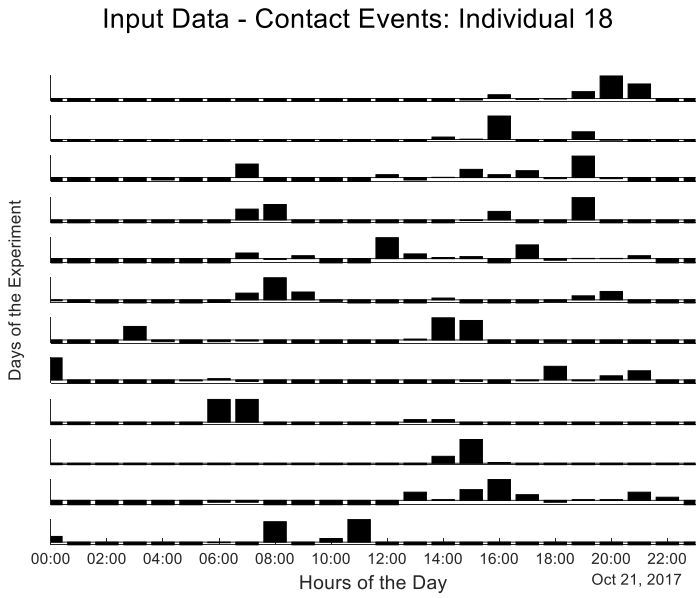

Clean Data: Individual 20
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 30, 2017

Appendix E – PCA Results

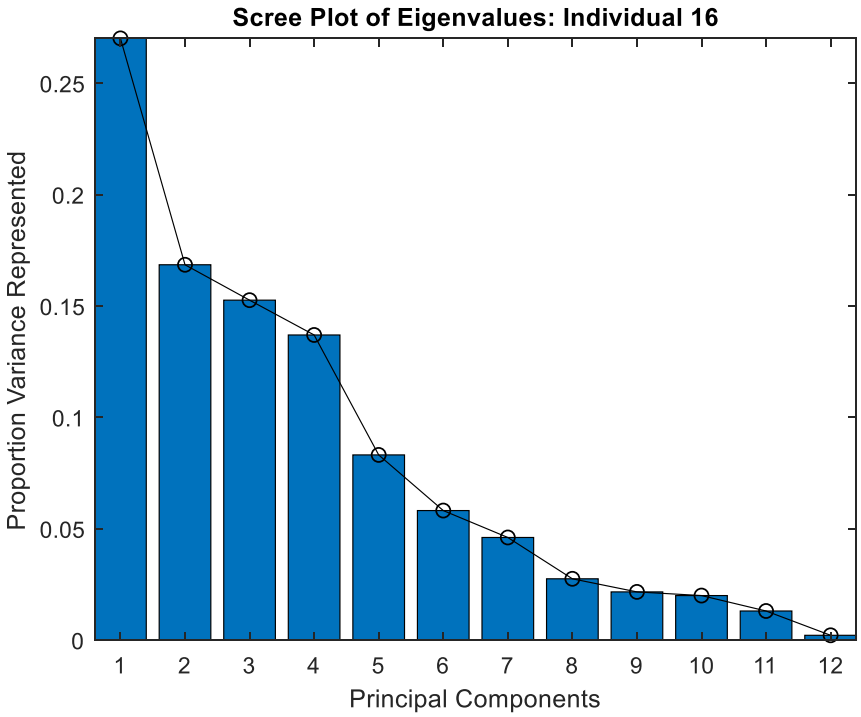
E1. Class 1

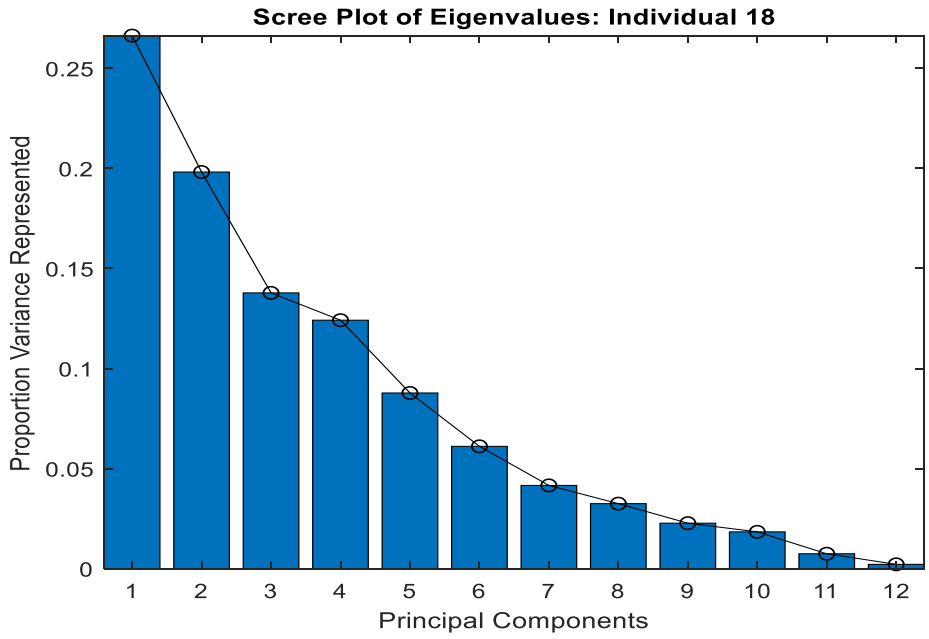
E1.1 Time Trends





E1.2 Scree Plots

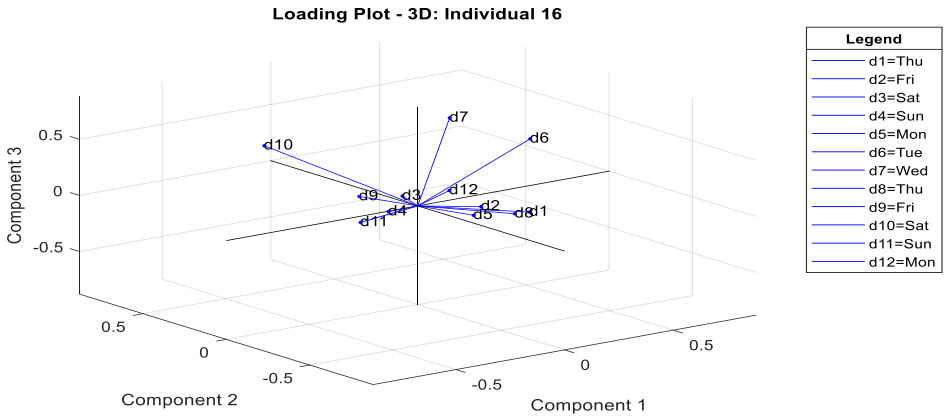


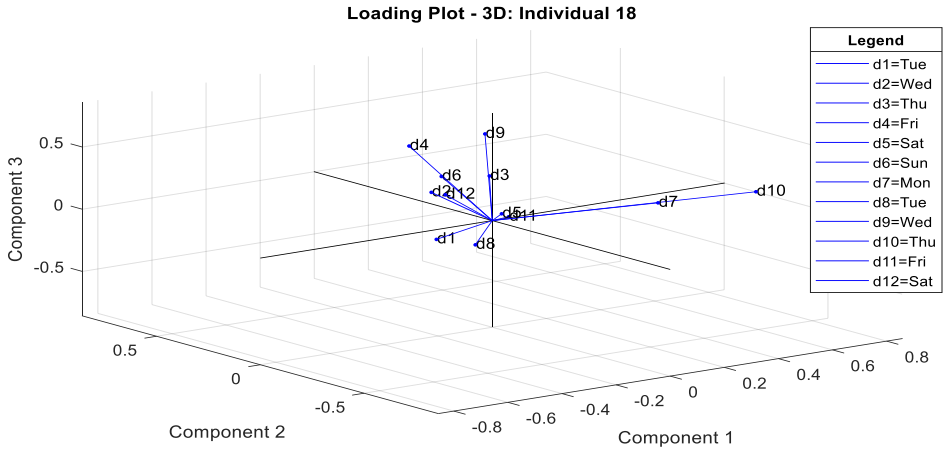


E1.3 Principal Component Loadings

PC Loadings - Individual 16				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,5517	0,0490	0,2715	**0,380**
Day 2	0,2119	-0,1044	0,0336	**0,057**
Day 3	0,0044	0,0976	-0,0411	**0,011**
Day 4	0,0027	0,1791	0,1345	**0,050**
Day 5	0,2687	0,0147	0,1849	**0,107**
Day 6	0,4545	-0,0820	-0,4759	**0,440**
Day 7	0,2314	0,1100	-0,6526	**0,492**
Day 8	0,4751	0,0368	0,2538	**0,292**
Day 9	-0,0158	0,3297	0,0615	**0,113**
Day 10	-0,0915	0,8005	-0,2018	**0,690**
Day 11	0,0243	0,3747	0,3277	**0,248**
Day 12	0,2889	0,1865	0,0501	**0,121**

PC Loadings - Individual 18				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,00723	-0,27835	-0,28107	**0,157**
Day 2	-0,29481	-0,67783	-0,18699	**0,581**
Day 3	-0,17008	-0,23523	0,191375	**0,121**
Day 4	-0,05975	-0,47849	0,359374	**0,362**
Day 5	-0,03085	0,004961	0,046154	**0,003**
Day 6	0,067987	-0,15632	0,306923	**0,123**
Day 7	-0,43398	0,23313	0,096979	**0,252**
Day 8	0,083297	0,027113	-0,15304	**0,031**
Day 9	0,025559	-0,00171	0,705103	**0,498**
Day 10	-0,77947	0,254614	0,07459	**0,678**
Day 11	-0,20289	-0,18549	-0,12553	**0,091**
Day 12	0,174059	0,000775	0,269462	**0,103**

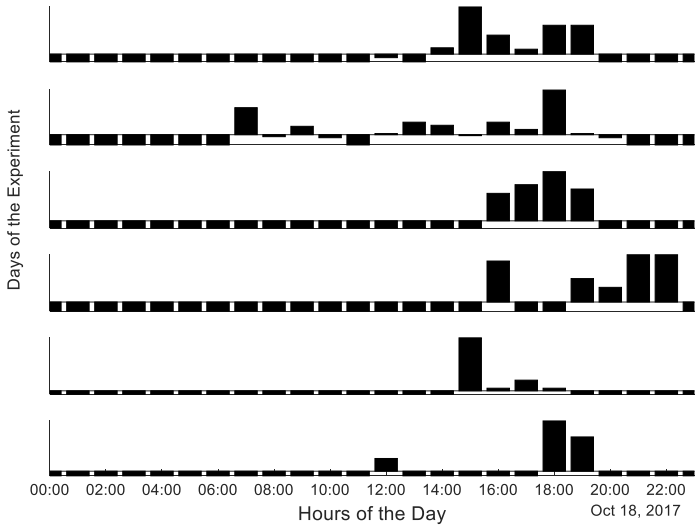




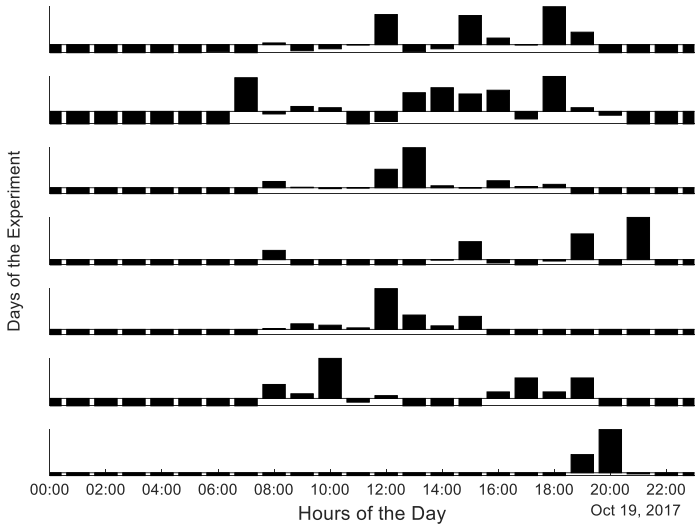
E2. Class 2

E2.1 Time Trends

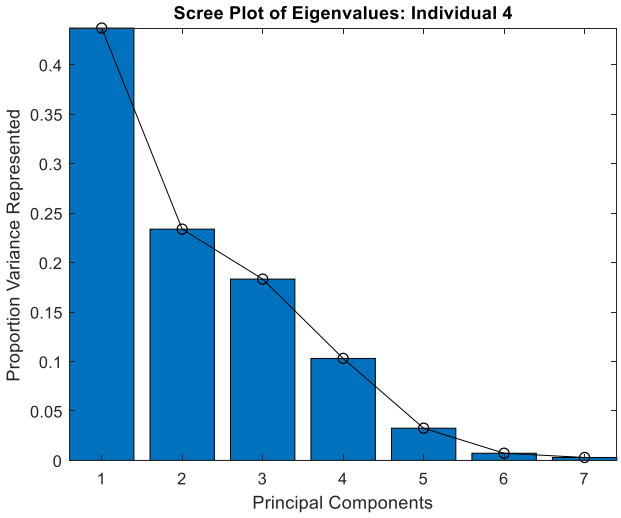
Input Data - Contact Events: Individual 13

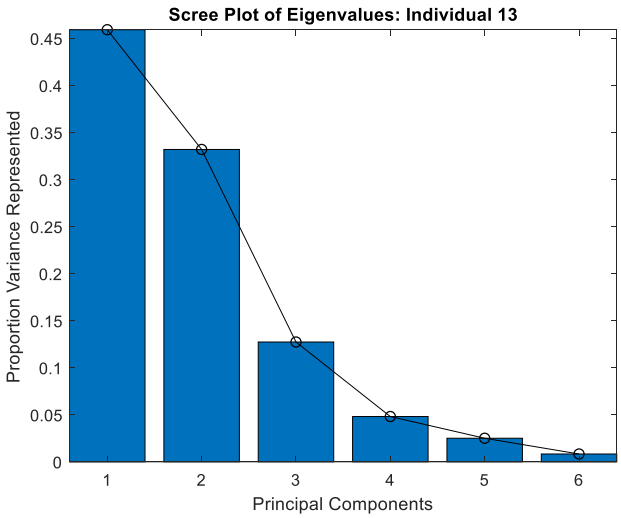


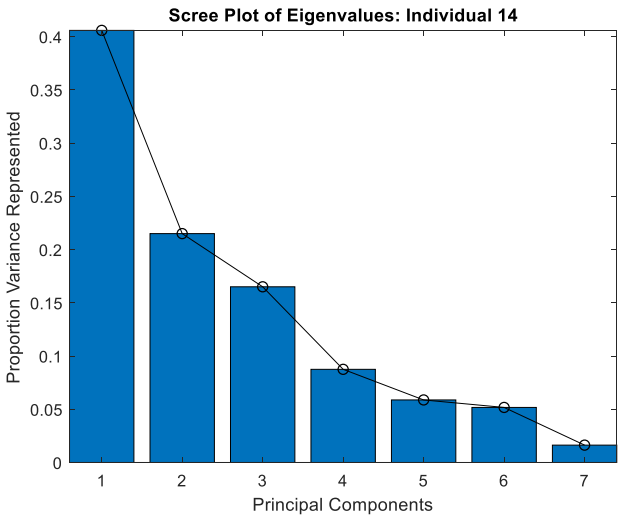
Input Data - Contact Events: Individual 14



E2.2 Scree Plot





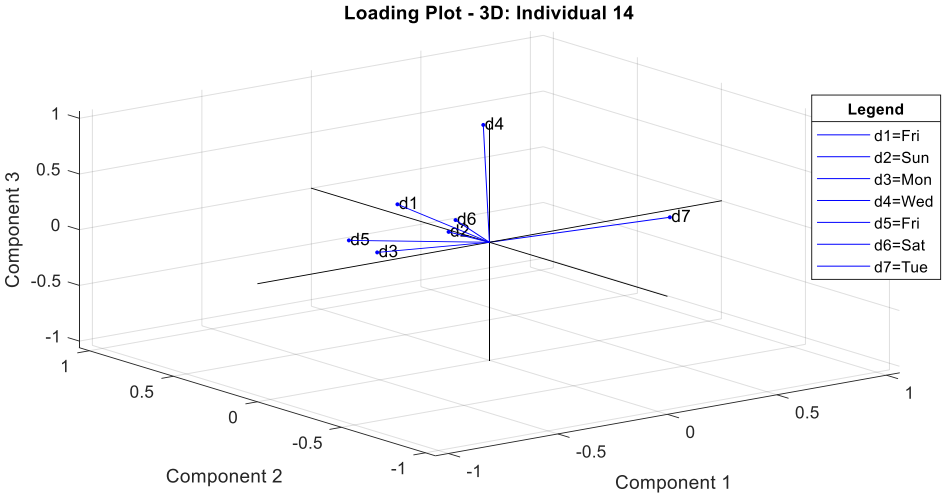


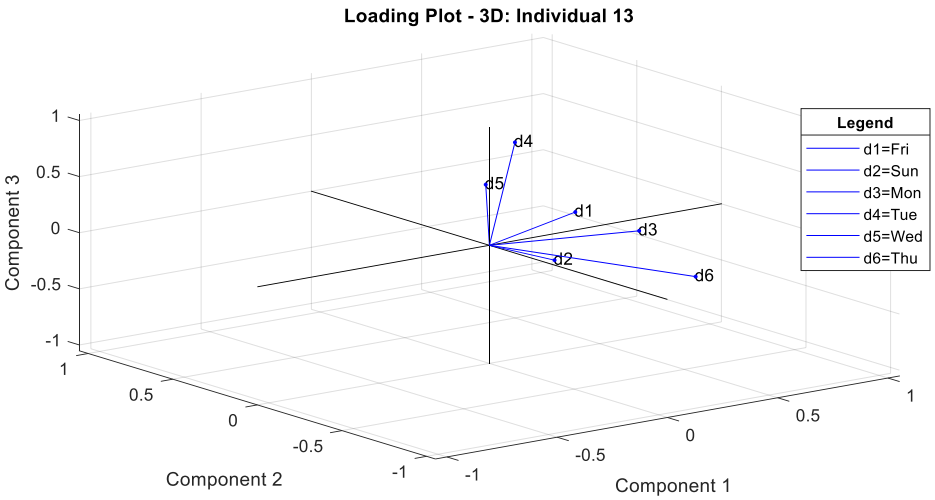
E2.3 Principal Component Loadings

PC Loadings - Individual 4				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,2770	0,0205	-0,2122	**0,122**
Day 2	0,2213	0,0459	0,2908	**0,136**
Day 3	0,3126	0,2270	0,0306	**0,150**
Day 4	0,5214	0,0991	0,5990	**0,640**
Day 5	0,4821	-0,1018	0,0614	**0,247**
Day 6	0,2958	0,6829	-0,5499	**0,856**
Day 7	0,4299	-0,6778	-0,4522	**0,849**

PC Loadings - Individual 13				Contribution
Variable	PC 1	PC 2	PC 2	
Day 1	0,4604	-0,0924	-0,0919	**0,229**
Day 2	0,1887	0,1393	0,1351	**0,073**
Day 3	0,4301	0,3245	-0,1266	**0,306**
Day 4	0,0313	0,1102	-0,9565	**0,928**
Day 5	0,5522	-0,7413	-0,0073	**0,854**
Day 6	0,5115	0,5524	0,2058	**0,609**

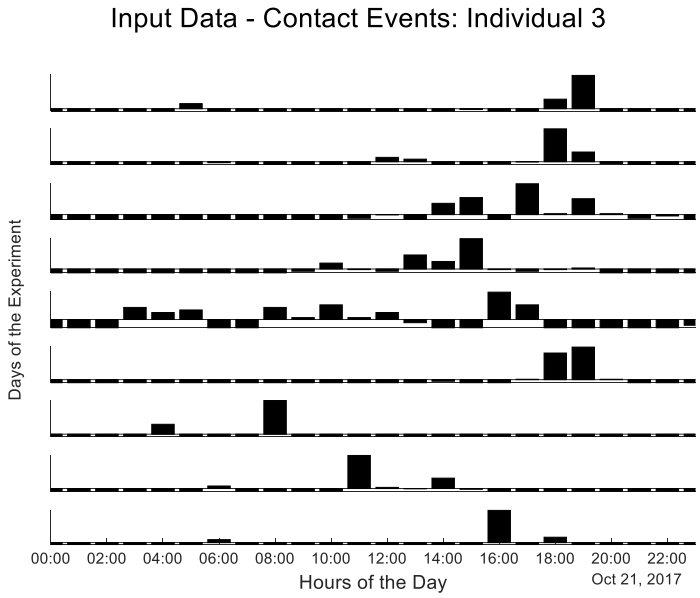
PC Loadings - Individual 14				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,0526	-0,4786	-0,1401	**0,251**
Day 2	-0,0298	-0,2039	-0,0103	**0,043**
Day 3	-0,1526	-0,4689	0,2522	**0,307**
Day 4	0,1367	-0,2144	-0,9072	**0,888**
Day 5	-0,1641	-0,6224	0,2105	**0,459**
Day 6	0,0051	-0,2080	-0,1025	**0,054**
Day 7	0,9630	-0,1813	0,1972	**0,999**

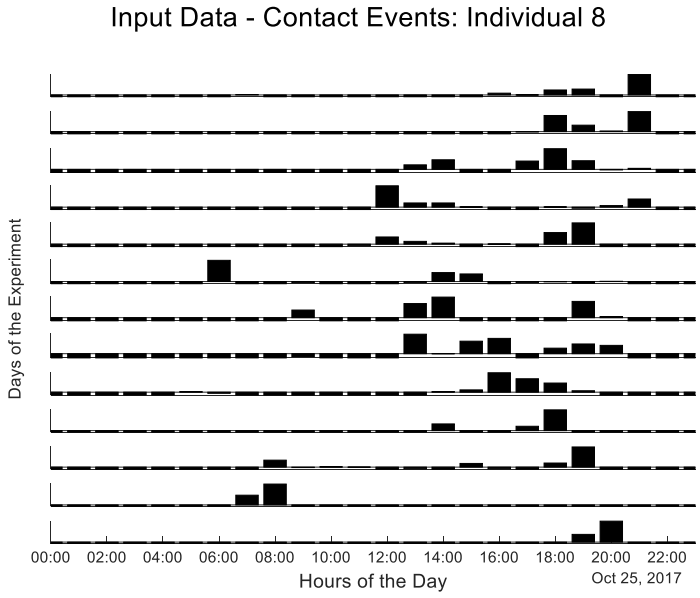

Loading Plot - 3D: Individual 14
Legend
d1=Fri
d2=Sun
d3=Mon
d4=Wed
d5=Fri
d6=Sat
d7=Tue
d1
d2
d3
d4
d5
d6
d7
Component 3
Component 2
Component 1


Loading Plot - 3D: Individual 13
Legend
d1=Fri
d2=Sun
d3=Mon
d4=Tue
d5=Wed
d6=Thu
d1
d2
d3
d4
d5
d6
Component 3
Component 2
Component 1

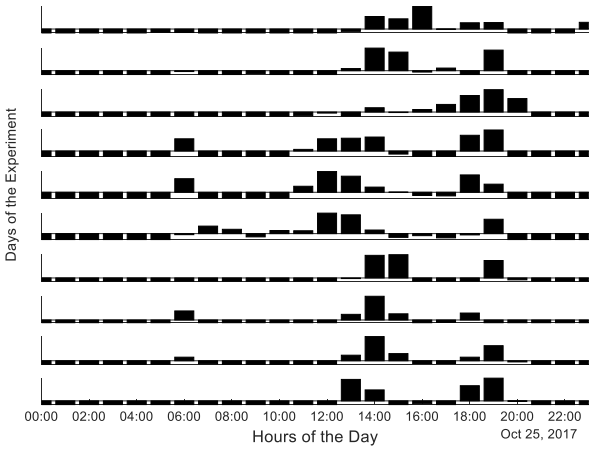
E3. Class 3

E3.1 Time Trends

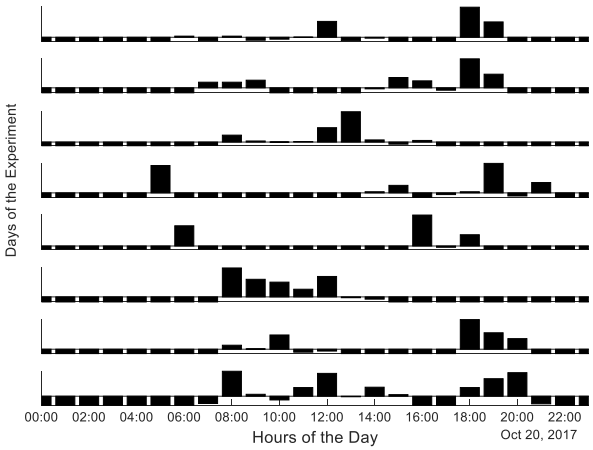




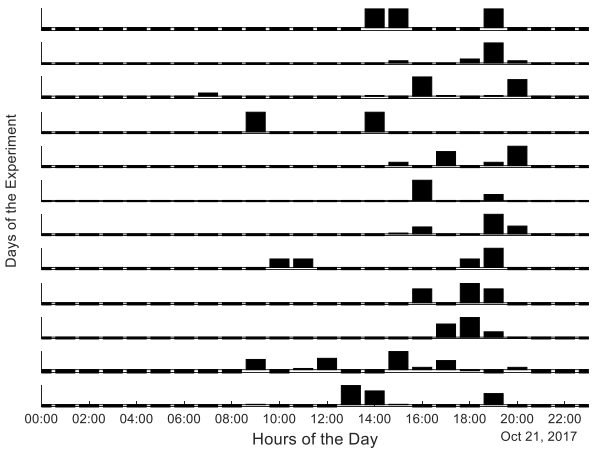
Input Data - Contact Events: Individual 9



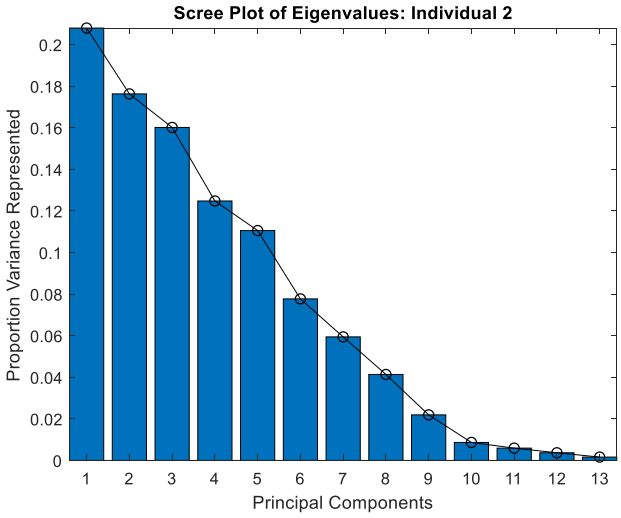
Input Data - Contact Events: Individual 12

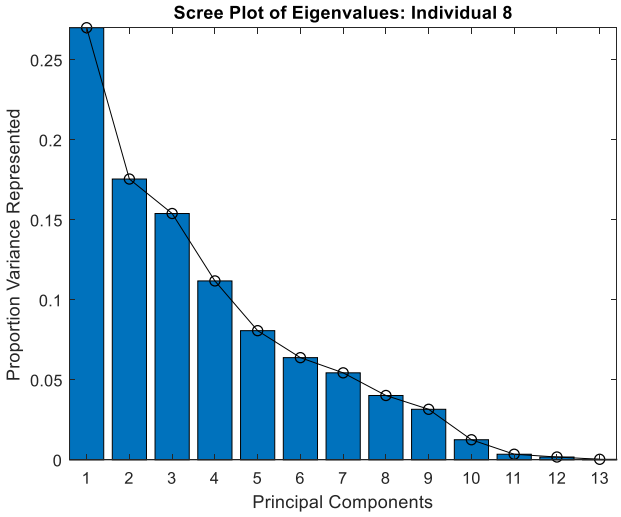


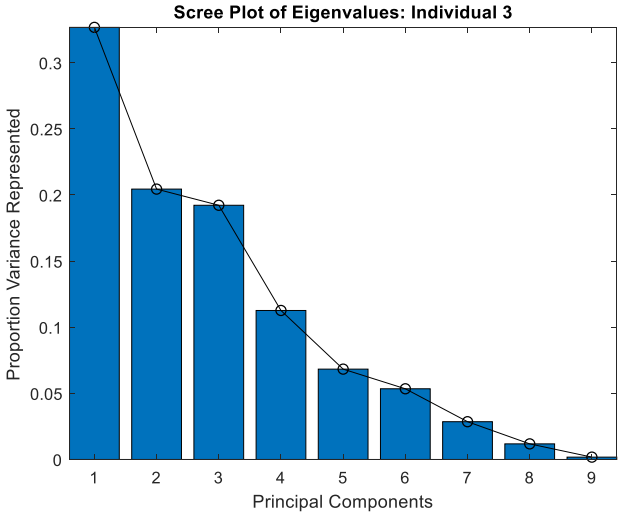
Input Data - Contact Events: Individual 20

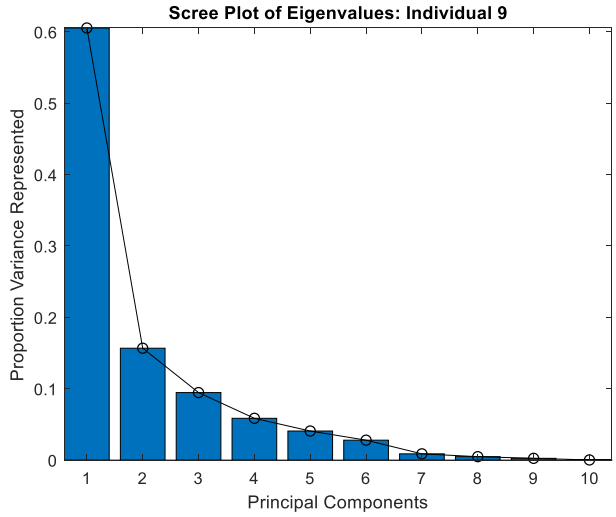


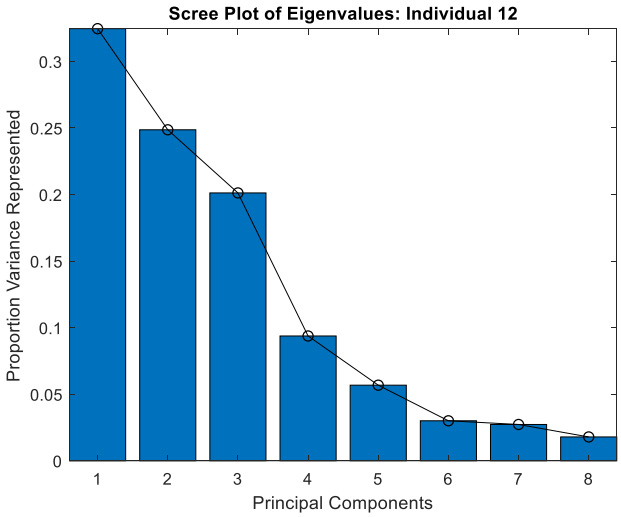
E3.2 Scree Plot

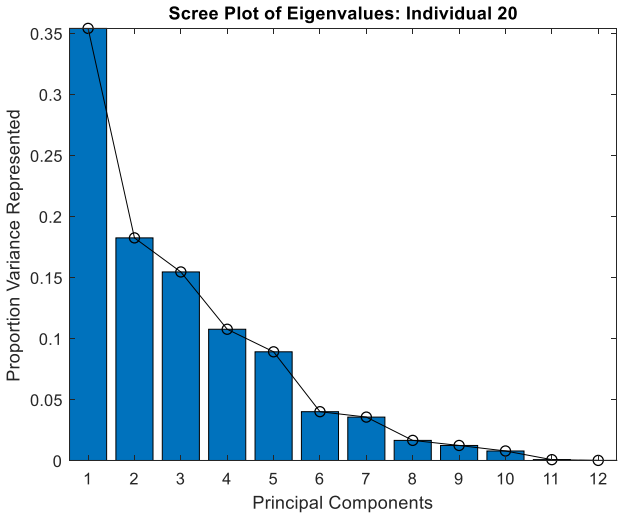







Scree Plot of Eigenvalues: Individual 9
Proportion Variance Represented
Principal Components
0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5
6
7
8
9
10


Scree Plot of Eigenvalues: Individual 12
Proportion Variance Represented
Principal Components
0
0.05
0.1
0.15
0.2
0.25
0.3
1
2
3
4
5
6
7
8


Scree Plot of Eigenvalues: Individual 20
Proportion Variance Represented
Principal Components
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
1
2
3
4
5
6
7
8
9
10
11
12

E3.3 Principal Component Loadings

PC Loadings - Individual 2				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,310	0,283	0,143	**0,196**
Day 2	-0,027	-0,399	0,162	**0,186**
Day 3	-0,493	0,428	0,045	**0,429**
Day 4	0,286	0,141	0,036	**0,103**
Day 5	0,213	-0,097	0,428	**0,238**
Day 6	0,143	0,199	0,282	**0,140**
Day 7	0,099	0,007	0,166	**0,037**
Day 8	-0,184	0,426	0,318	**0,316**
Day 9	0,057	0,058	0,144	**0,027**
Day 10	0,458	0,399	-0,409	**0,536**
Day 11	0,191	0,348	-0,298	**0,247**
Day 12	0,466	0,125	0,414	**0,404**
Day 13	0,042	0,167	0,333	**0,141**

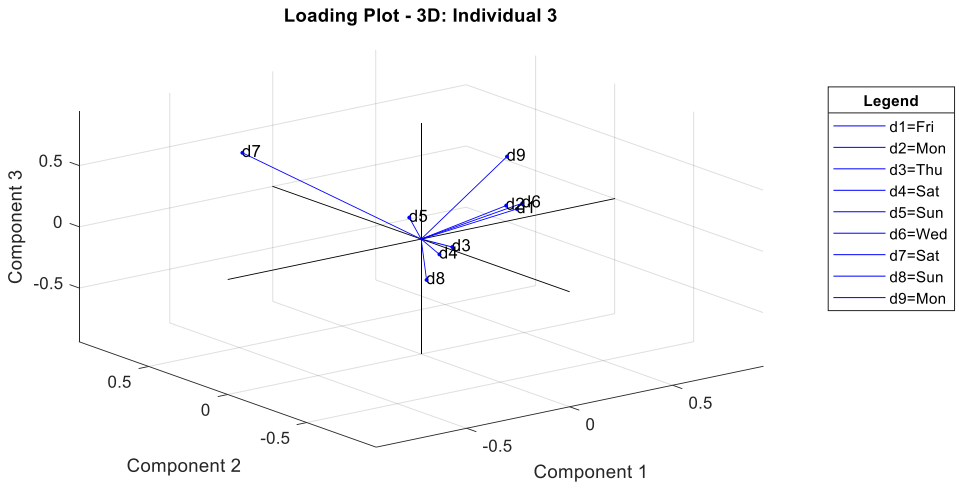
PC Loadings - Individual 3				Contribution
Variables	PC 1	PC 2	PC 3	
Day 1	0,5964	-0,1722	-0,0428	**0,387**
Day 2	0,4475	-0,0454	0,0954	**0,211**
Day 3	0,1551	-0,0055	-0,1220	**0,039**
Day 4	0,0312	0,0746	-0,1017	**0,017**
Day 5	-0,0617	0,0050	0,1993	**0,044**
Day 6	0,5967	-0,1378	0,0200	**0,375**
Day 7	-0,2129	-0,8539	0,3883	**0,925**
Day 8	-0,0870	0,1474	-0,2357	**0,085**
Day 9	0,0797	0,4392	0,8470	**0,917**

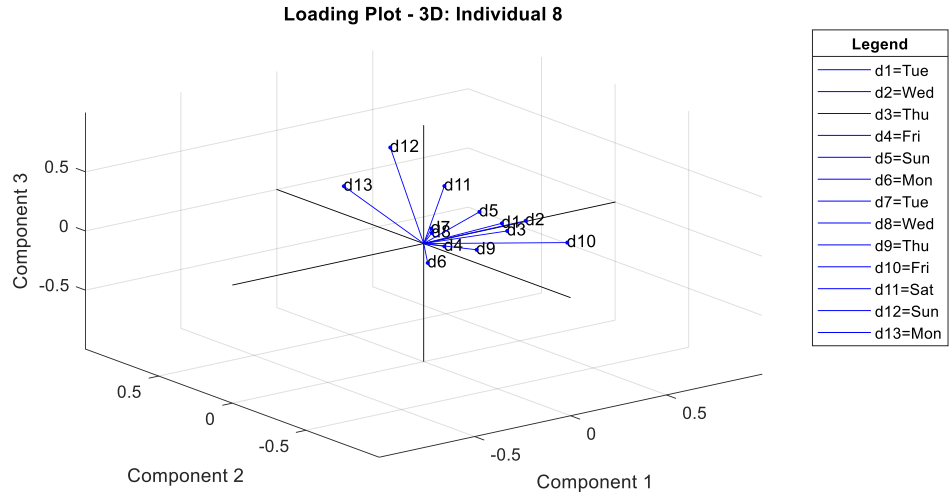
PC Loadings - Individual 8				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,3162	-0,1187	0,1144	**0,127**
Day 2	0,4249	-0,1389	0,1069	**0,211**
Day 3	0,3559	-0,1032	0,0286	**0,138**
Day 4	0,1160	0,0108	-0,0690	**0,018**
Day 5	0,3593	0,0897	0,1002	**0,147**
Day 6	0,0129	-0,0105	-0,1630	**0,027**
Day 7	0,1811	0,1843	-0,0216	**0,067**
Day 8	0,1569	0,1490	-0,0334	**0,048**
Day 9	0,1680	-0,1403	-0,0472	**0,050**
Day 10	0,4613	-0,3742	0,0174	**0,353**
Day 11	0,2369	0,1682	0,3259	**0,191**
Day 12	-0,2103	-0,0469	0,9069	**0,869**
Day 13	0,2265	0,8375	0,0215	**0,753**

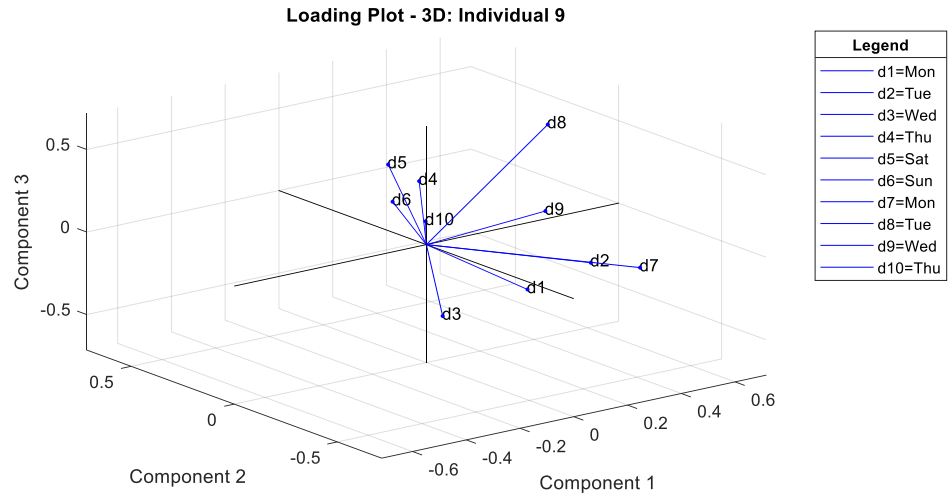
PC Loadings - Individual 9				Contribution
Variable	PC 1	PC 2	Pc 3	
Day 1	0,2144	-0,2109	0,2505	**0,153**
Day 2	0,4001	-0,2760	0,1238	**0,252**
Day 3	0,2228	0,2114	0,6077	**0,464**
Day 4	0,2636	0,3786	-0,1183	**0,227**
Day 5	0,1739	0,4120	-0,2351	**0,255**
Day 6	0,1211	0,3216	-0,0701	**0,123**
Day 7	0,4615	-0,4347	0,1024	**0,413**
Day 8	0,3661	-0,1116	-0,6500	**0,569**
Day 9	0,3991	-0,0570	-0,0892	**0,170**
Day 10	0,3515	0,4620	0,1940	**0,375**

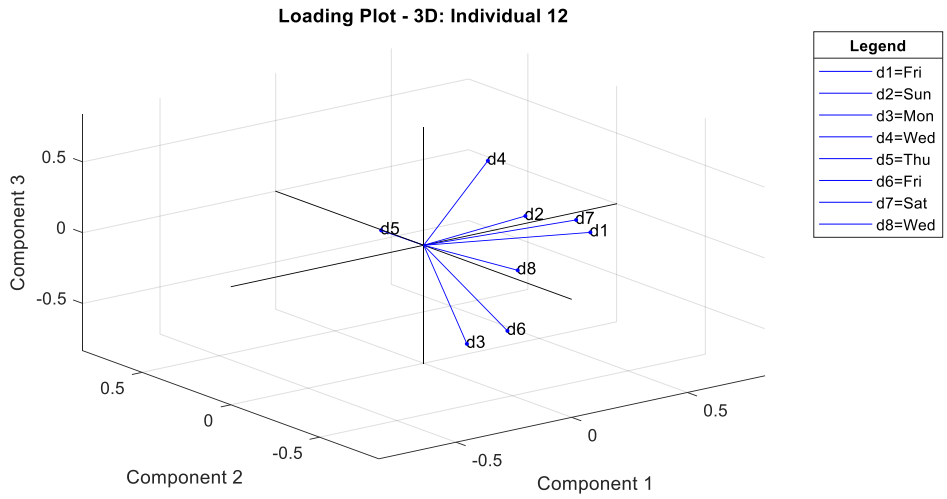
PC Loadings - Individual 12				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,5704	0,1959	0,0195	**0,364**
Day 2	0,4323	0,0093	-0,0590	**0,190**
Day 3	-0,0328	0,2859	0,5539	**0,390**
Day 4	0,1939	0,1090	-0,5793	**0,385**
Day 5	0,3978	-0,7581	0,3812	**0,878**
Day 6	0,0475	0,4100	0,4337	**0,358**
Day 7	0,5052	0,2004	-0,0932	**0,304**
Day 8	0,1829	0,2914	0,1077	**0,130**

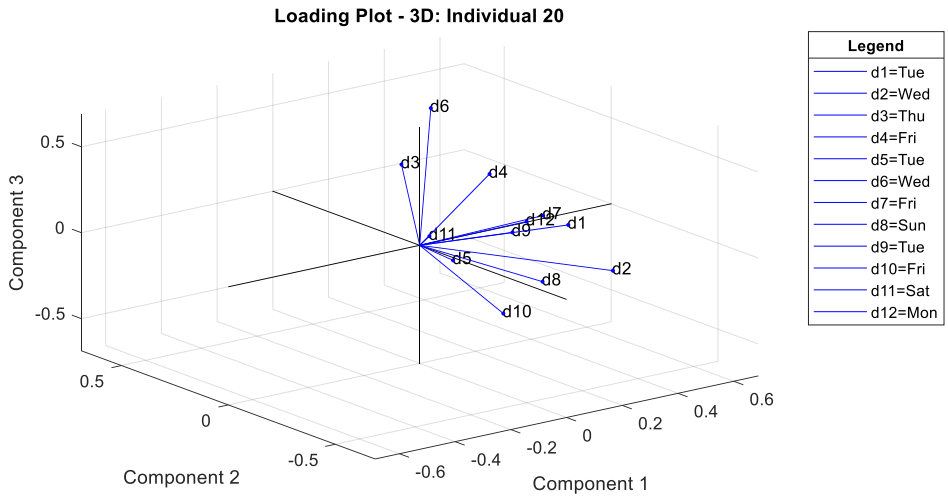
PC Loadings - Individual 20				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,2246	0,4022	-0,2239	**0,262**
Day 2	0,4626	0,3009	0,1722	**0,334**
Day 3	0,2141	-0,3644	-0,2281	**0,231**
Day 4	-0,0632	0,4092	-0,6246	**0,562**
Day 5	0,1324	-0,0163	0,1414	**0,038**
Day 6	0,4406	-0,5222	-0,4040	**0,630**
Day 7	0,4266	0,0216	-0,0334	**0,184**
Day 8	0,2845	0,2040	0,2181	**0,170**
Day 9	0,3889	-0,0731	0,0951	**0,166**
Day 10	0,2088	0,1190	0,4156	**0,230**
Day 11	0,0160	0,0248	-0,0589	**0,004**
Day 12	0,1267	0,3352	-0,2469	**0,189**


Loading Plot - 3D: Individual 3
Component 3
Component 2
Component 1
d7
d9
d5
d3
d4
d8
Legend
d1=Fri
d2=Mon
d3=Thu
d4=Sat
d5=Sun
d6=Wed
d7=Sat
d8=Sun
d9=Mon


Loading Plot - 3D: Individual 8
Component 3
Component 2
Component 1
d12
d13
d11
d5
d10
d6
Legend
d1=Tue
d2=Wed
d3=Thu
d4=Fri
d5=Sun
d6=Mon
d7=Tue
d8=Wed
d9=Thu
d10=Fri
d11=Sat
d12=Sun
d13=Mon


Loading Plot - 3D: Individual 9
Component 3
Component 2
Component 1
d8
d5
d4
d6
d10
d9
d2
d7
d1
d3
Legend
d1=Mon
d2=Tue
d3=Wed
d4=Thu
d5=Sat
d6=Sun
d7=Mon
d8=Tue
d9=Wed
d10=Thu

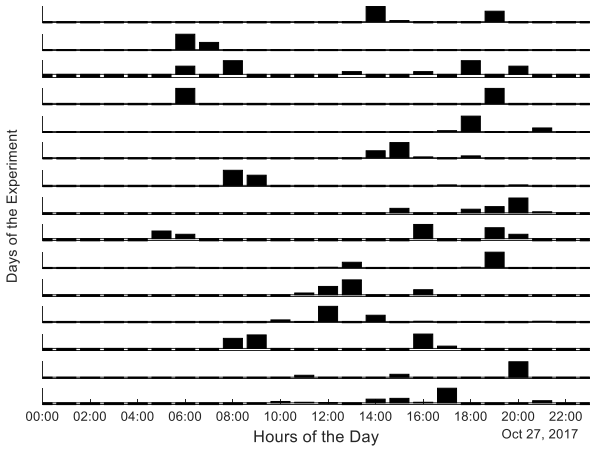




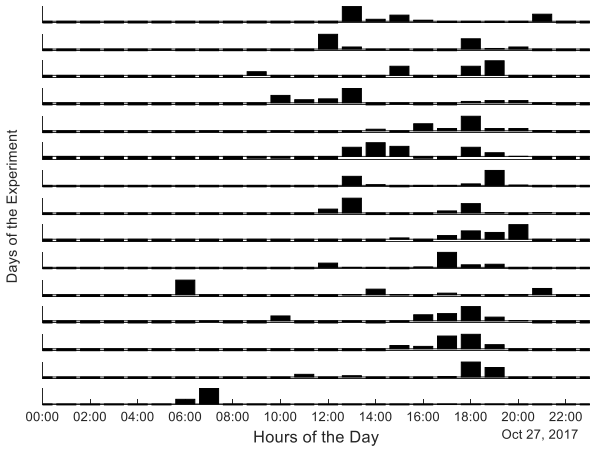
E4. Class 4

E4.1 Time Trends

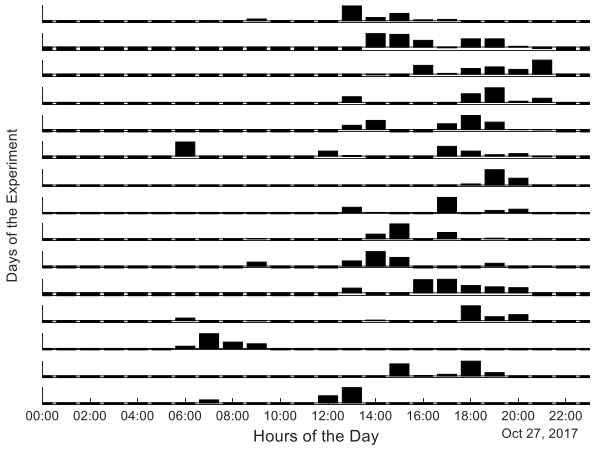
Input Data - Contact Events: Individual 5

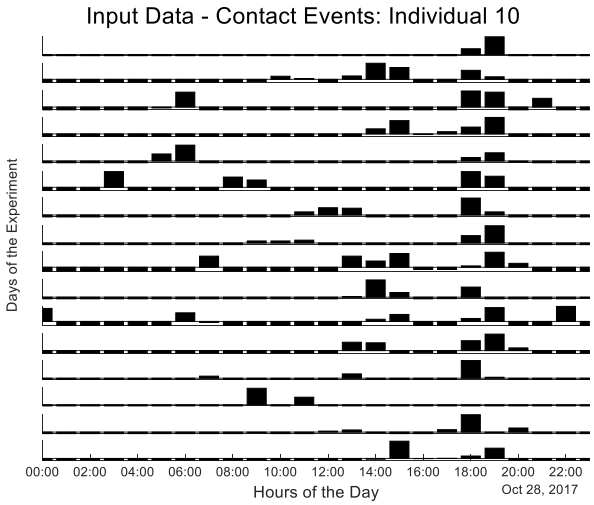


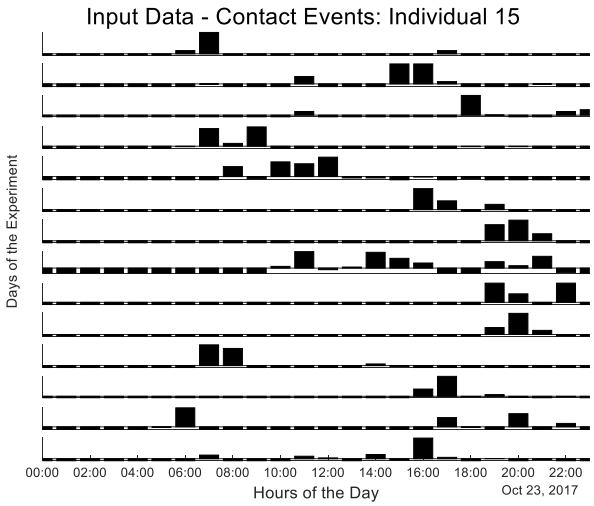
Input Data - Contact Events: Individual 6

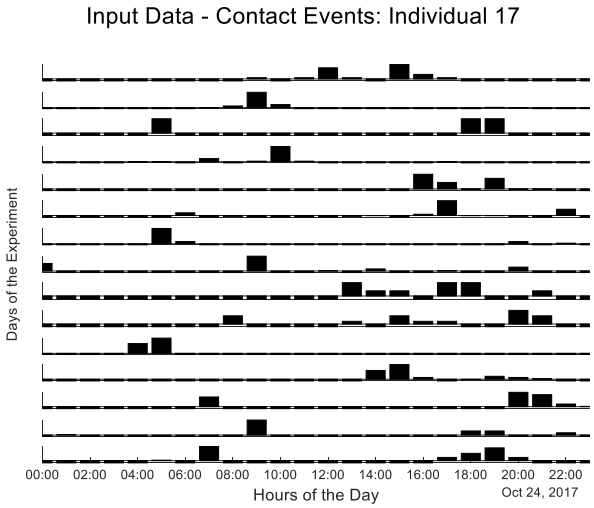


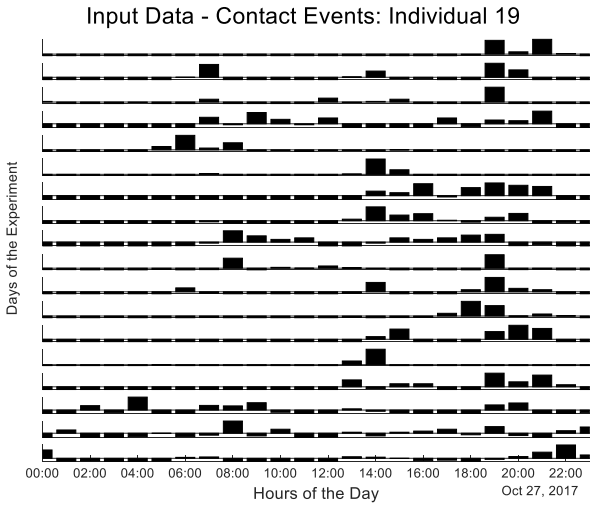
Input Data - Contact Events: Individual 7



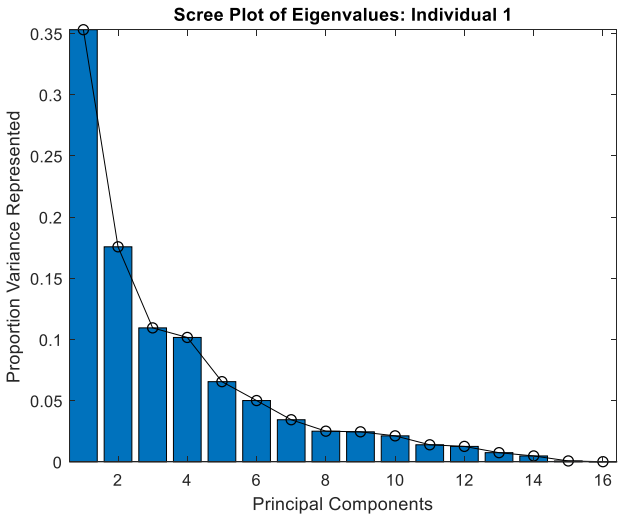

Input Data - Contact Events: Individual 10
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 28, 2017

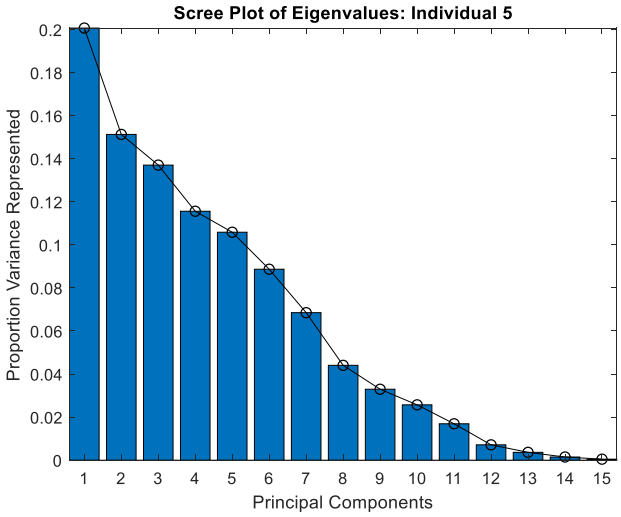

Input Data - Contact Events: Individual 15
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 23, 2017

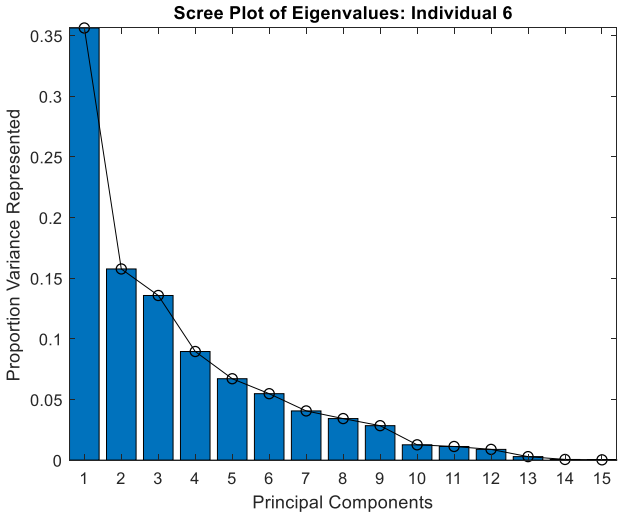

Input Data - Contact Events: Individual 17
Days of the Experiment
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
Hours of the Day
Oct 24, 2017

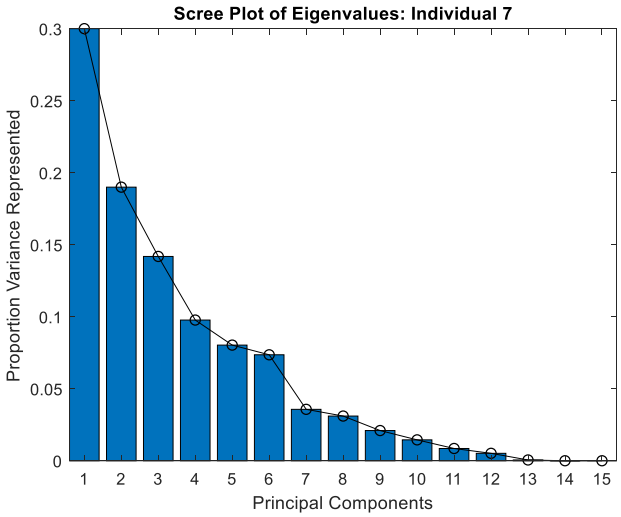


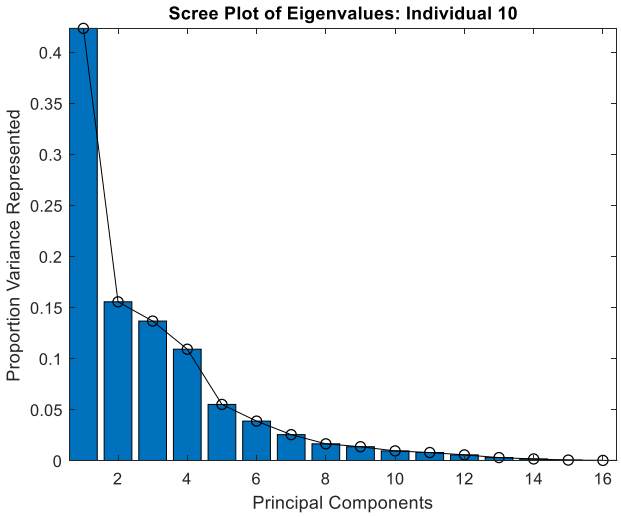
E4.2 Scree Plots

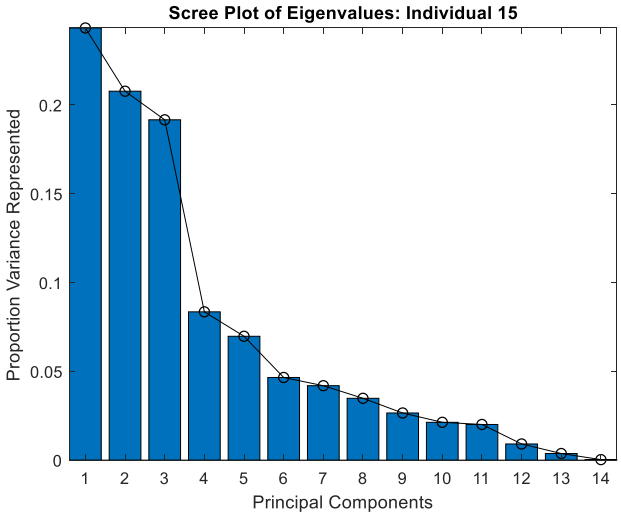


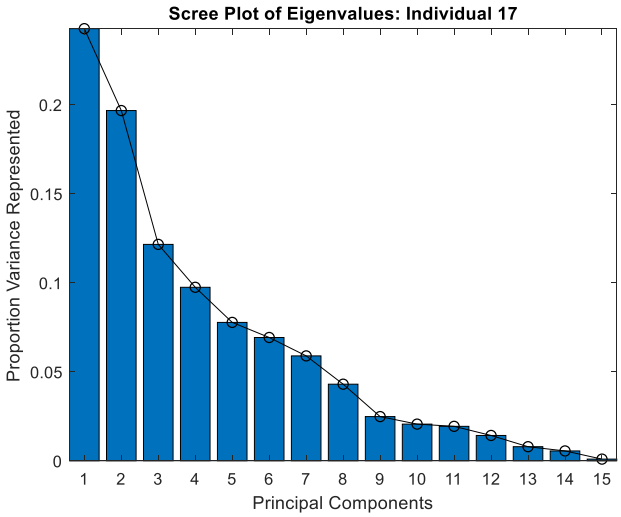


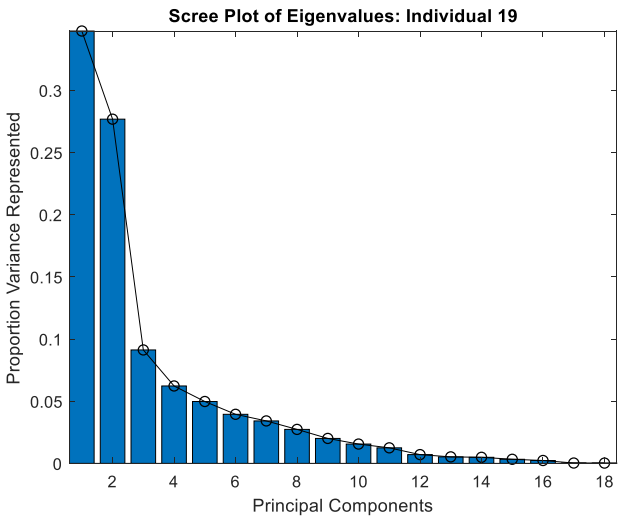

Scree Plot of Eigenvalues: Individual 6
Proportion Variance Represented
Principal Components
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


Scree Plot of Eigenvalues: Individual 7
Proportion Variance Represented
Principal Components
0
0.05
0.1
0.15
0.2
0.25
0.3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


Scree Plot of Eigenvalues: Individual 10
Proportion Variance Represented
Principal Components
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
2 4 6 8 10 12 14 16


Scree Plot of Eigenvalues: Individual 15
Proportion Variance Represented
0
0.05
0.1
0.15
0.2
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Principal Components


Scree Plot of Eigenvalues: Individual 17
Proportion Variance Represented
0
0.05
0.1
0.15
0.2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Principal Components


Scree Plot of Eigenvalues: Individual 19
Proportion Variance Represented
0
0.05
0.1
0.15
0.2
0.25
0.3
2 4 6 8 10 12 14 16 18
Principal Components

E4.3 Principal Component Loadings

PC Loadings - Individual 1				Contribution
Variale	PC 1	PC 2	PC 3	
Day 1	0,5291	-0,3376	0,0654	**0,398**
Day 2	0,3753	-0,2085	-0,1165	**0,198**
Day 3	0,2020	-0,1929	-0,3603	**0,208**
Day 4	0,4700	-0,0393	0,4327	**0,410**
Day 5	0,1924	0,1128	-0,3065	**0,144**
Day 6	0,2469	0,3509	0,3221	**0,288**
Day 7	0,0644	0,2771	-0,0336	**0,082**
Day 8	-0,0477	-0,0289	-0,1310	**0,020**
Day 9	0,2256	0,5426	-0,3306	**0,455**
Day 10	0,2519	0,0323	-0,2156	**0,111**
Day 11	0,2094	-0,1457	0,1323	**0,083**
Day 12	0,1766	0,1316	-0,4185	**0,224**
Day 13	0,0817	0,1665	0,1817	**0,067**
Day 14	-0,0443	-0,0971	-0,1081	**0,023**
Day 15	0,0891	0,0704	-0,0996	**0,023**
Day 16	0,0849	0,4628	0,2147	**0,268**

PC Loadings - Individual 5				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,3193	-0,1960	0,4472	**0,340**
Day 2	-0,3802	0,3093	-0,4200	**0,417**
Day 3	-0,0222	-0,0851	-0,2887	**0,091**
Day 4	-0,6353	0,0848	-0,1349	**0,429**
Day 5	0,0909	-0,1872	-0,1381	**0,062**
Day 6	0,0169	-0,2928	0,2448	**0,146**
Day 7	0,1566	0,1155	-0,2553	**0,103**
Day 8	-0,1169	-0,4966	-0,1607	**0,286**
Day 9	-0,2295	-0,0675	-0,0610	**0,061**
Day 10	-0,4784	-0,0872	0,1926	**0,274**
Day 11	0,0162	0,0837	0,1878	**0,043**
Day 12	0,0327	0,0235	0,3691	**0,138**
Day 13	0,1202	0,1349	-0,1300	**0,050**
Day 14	0,0281	-0,6547	-0,3267	**0,536**
Day 15	0,0671	-0,0645	0,1270	**0,025**

PC Loadings - Individual 6				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,0944	-0,2501	0,4755	**0,298**
Day 2	-0,2467	-0,0056	0,0266	**0,062**
Day 3	-0,3079	0,0735	-0,0158	**0,100**
Day 4	-0,1404	-0,1794	0,3220	**0,156**
Day 5	-0,2785	0,1340	-0,1348	**0,114**
Day 6	-0,1964	-0,0792	0,2090	**0,089**
Day 7	-0,3116	-0,0828	0,3385	**0,219**
Day 8	-0,2944	-0,1120	0,3459	**0,219**
Day 9	-0,2669	0,1024	-0,1442	**0,103**
Day 10	-0,2329	0,1294	-0,2314	**0,125**
Day 11	0,0353	0,1021	0,1519	**0,035**
Day 12	-0,2671	0,1456	-0,1587	**0,118**
Day 13	-0,3339	0,1868	-0,2397	**0,204**
Day 14	-0,4216	0,1308	-0,0198	**0,195**
Day 15	0,1688	0,8633	0,4390	**0,967**

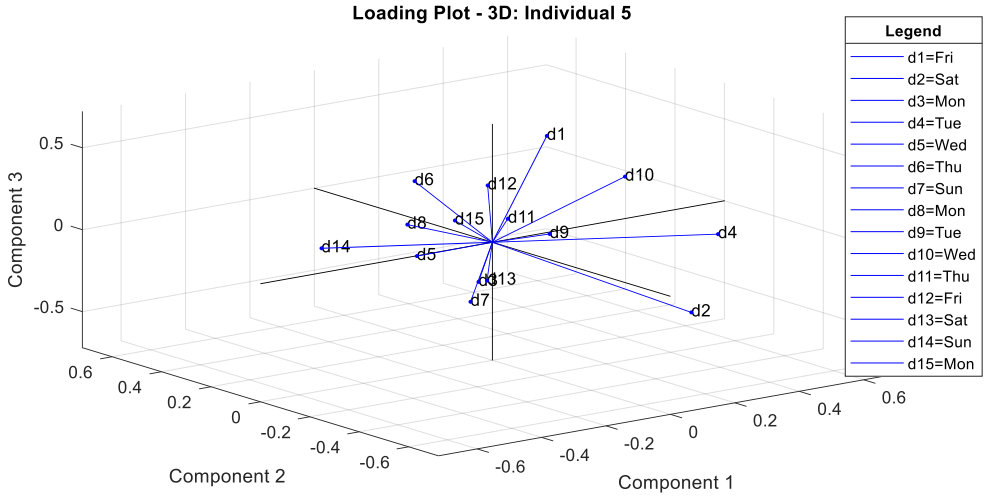
PC Loadings - Individual 7				Contribution
Variable	PC 1	PC 2	PC 3	
DaY 1	0,1051	0,5205	0,0078	**0,282**
Day 2	0,2512	0,0504	-0,2886	**0,149**
DaY 3	0,2138	-0,1445	0,0411	**0,068**
DaY 4	0,3737	-0,0742	0,2372	**0,201**
DaY 5	0,3135	0,0566	0,0118	**0,102**
DaY 6	0,1328	0,0136	0,1286	**0,034**
DaY 7	0,4660	-0,3097	0,2776	**0,390**
DaY 8	0,2179	0,2796	0,1982	**0,165**
DaY 9	0,1695	0,2770	-0,4967	**0,352**
DaY 10	0,1471	0,2969	-0,2693	**0,182**
DaY 11	0,2372	0,0648	0,0869	**0,068**
DaY 12	0,3383	-0,2069	0,0239	**0,158**
DaY 13	-0,1585	-0,0056	0,1066	**0,037**
DaY 14	0,3372	0,0404	-0,3305	**0,225**
DaY 15	0,0083	0,5559	0,5270	**0,587**

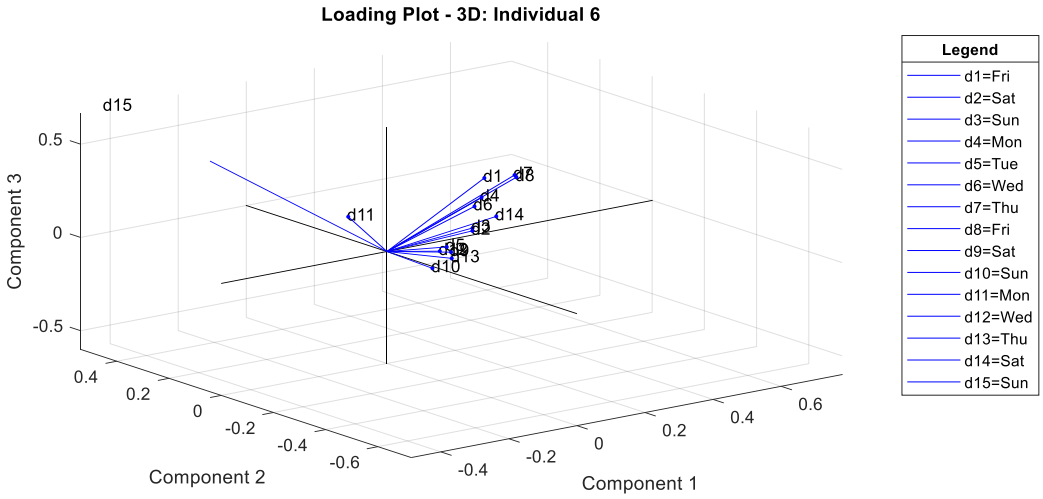
PC Loadings - Individual 10				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,4897	0,3817	-0,0979	**0,395**
Day 2	0,1506	-0,1031	0,1510	**0,056**
Day 3	0,2517	0,0429	-0,0434	**0,067**
Day 4	0,2608	0,1747	0,1372	**0,117**
Day 5	0,1641	0,1770	-0,0057	**0,058**
Day 6	0,1760	-0,0325	-0,2072	**0,075**
Day 7	0,2410	-0,3101	-0,1127	**0,167**
Day 8	0,3241	0,2191	-0,2184	**0,201**
Day 9	0,1428	0,0779	0,1391	**0,046**
Day 10	0,1799	-0,3029	0,2126	**0,169**
Day 11	0,1140	0,1531	0,0969	**0,046**
Day 12	0,2773	-0,0006	0,0275	**0,078**
Day 13	0,3444	-0,4979	-0,0873	**0,374**
Day 14	-0,0692	0,0631	-0,8449	**0,723**
Day 15	0,2522	-0,4336	-0,0711	**0,257**
Day 16	0,2356	0,2715	0,2056	**0,171**

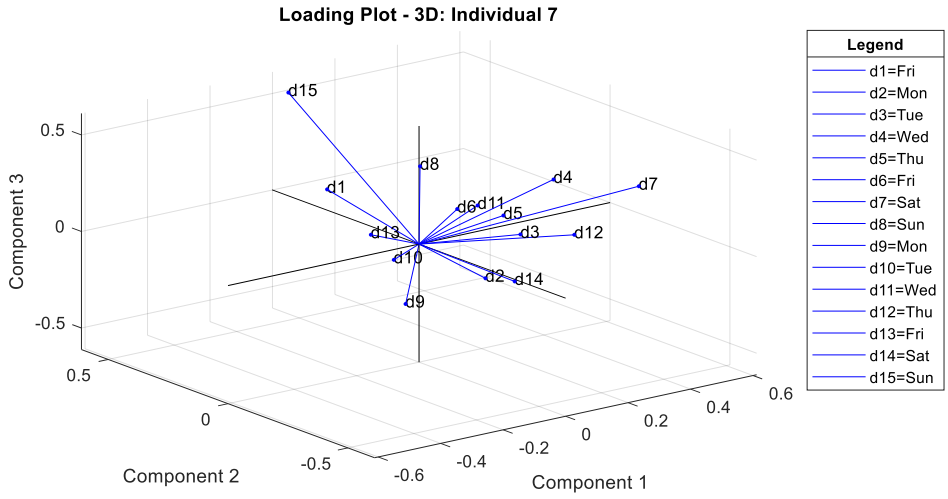
PC Loadings - Individual 15				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,4933	-0,0344	0,4998	**0,4942**
Day 2	0,0224	0,3584	0,0017	**0,1289**
Day 3	-0,0380	-0,0629	-0,1411	**0,0253**
Day 4	0,3269	-0,1309	0,2713	**0,1976**
Day 5	0,0467	-0,0079	-0,1220	**0,0171**
Day 6	-0,0713	0,5909	0,1572	**0,3789**
Day 7	-0,3965	-0,1236	0,4032	**0,3350**
Day 8	-0,0800	0,0589	0,0008	**0,0099**
Day 9	-0,3066	-0,0858	0,2602	**0,1690**
Day 10	-0,4262	-0,1622	0,4480	**0,4086**
Day 11	0,4219	-0,1198	0,3213	**0,2955**
Day 12	-0,0388	0,5216	0,1728	**0,3034**
Day 13	-0,1363	0,0021	0,2052	**0,0607**
Day 14	0,0578	0,4015	0,1049	**0,1756**

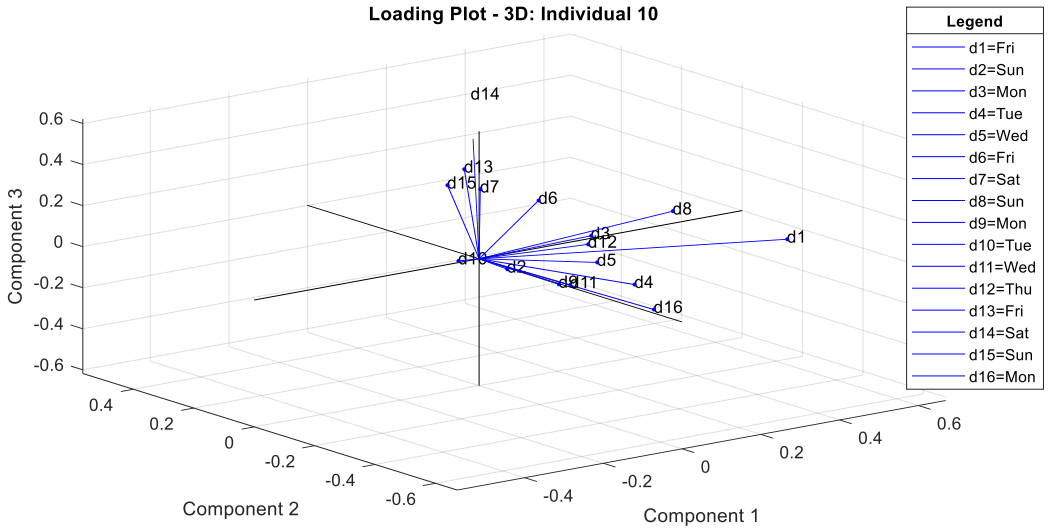
PC Loadings - Individual 17				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	-0,0873	-0,0849	0,0960	**0,024**
Day 2	-0,2739	0,6224	0,0031	**0,462**
Day 3	0,3214	0,1562	0,4334	**0,316**
Day 4	-0,0681	0,1573	-0,4894	**0,269**
Day 5	-0,0240	-0,0923	0,4730	**0,233**
Day 6	-0,0307	-0,1104	0,2407	**0,071**
Day 7	0,5884	0,2173	-0,0146	**0,394**
Day 8	-0,1994	0,3851	0,0570	**0,191**
Day 9	-0,0341	-0,1160	0,2204	**0,063**
Day 10	-0,0461	-0,1226	0,0559	**0,020**
Day 11	0,6152	0,2266	-0,1079	**0,441**
Day 12	-0,0500	-0,1296	0,1903	**0,056**
Day 13	-0,0238	-0,0816	-0,0620	**0,011**
Day 14	-0,1869	0,4865	0,2901	**0,356**
Day 15	0,0466	0,0076	0,3012	**0,093**

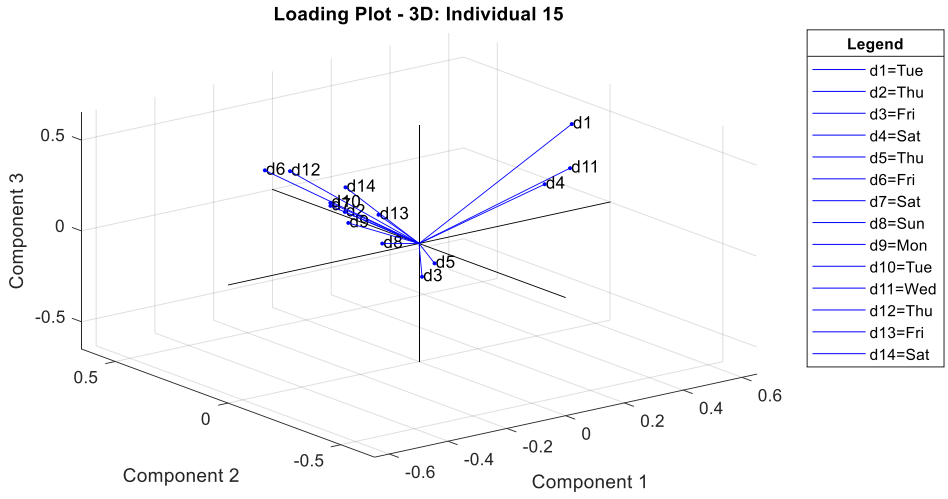
PC Loadings - Individual 19				Contribution
Variable	PC 1	PC 2	PC 3	
Day 1	0,1632	0,4809	0,2392	**0,315**
Day 2	0,2279	0,1747	-0,1796	**0,115**
Day 3	0,1839	0,3025	-0,2911	**0,210**
Day 4	0,0001	0,1300	0,0839	**0,024**
Day 5	-0,0603	-0,0487	-0,5149	**0,271**
Day 6	0,5220	-0,3139	0,0044	**0,371**
Day 7	0,1531	0,1920	0,1816	**0,093**
Day 8	0,2630	-0,0309	0,0941	**0,079**
Day 9	0,0209	0,0850	-0,1776	**0,039**
Day 10	0,1096	0,2734	-0,5001	**0,337**
Day 11	0,2940	0,2025	-0,1433	**0,148**
Day 12	0,0815	0,3043	-0,0176	**0,100**
Day 13	0,1896	0,2258	0,3169	**0,187**
Day 14	0,5990	-0,3837	-0,0305	**0,507**
Day 15	0,1317	0,2665	0,1893	**0,124**
Day 16	0,0190	0,0584	-0,0867	**0,011**
Day 17	-0,0034	0,0506	-0,1604	**0,028**
Day 18	0,0305	0,0177	0,1971	**0,040**

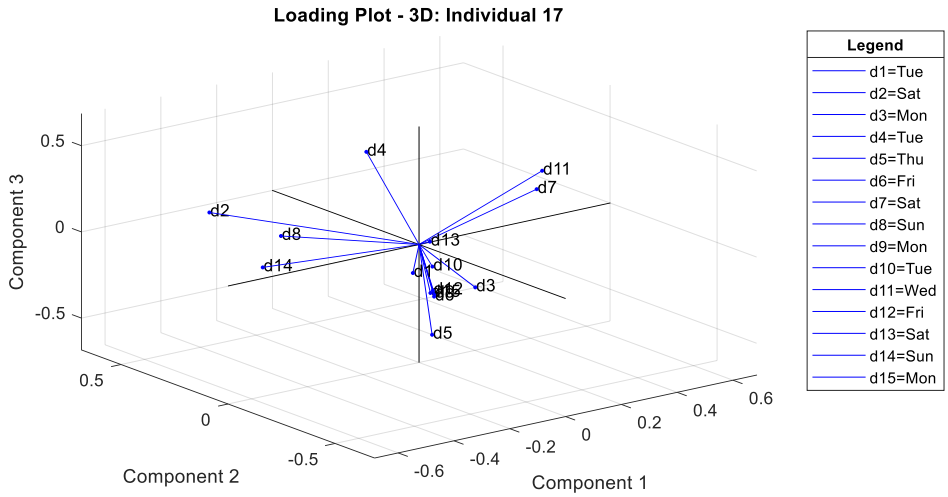

Loading Plot - 3D: Individual 5
Legend
d1=Fri
d2=Sat
d3=Mon
d4=Tue
d5=Wed
d6=Thu
d7=Sun
d8=Mon
d9=Tue
d10=Wed
d11=Thu
d12=Fri
d13=Sat
d14=Sun
d15=Mon
Component 3
Component 2
Component 1


Loading Plot - 3D: Individual 6
Legend
d1=Fri
d2=Sat
d3=Sun
d4=Mon
d5=Tue
d6=Wed
d7=Thu
d8=Fri
d9=Sat
d10=Sun
d11=Mon
d12=Wed
d13=Thu
d14=Sat
d15=Sun
Component 3
Component 2
Component 1


Loading Plot - 3D: Individual 7
Legend
d1=Fri
d2=Mon
d3=Tue
d4=Wed
d5=Thu
d6=Fri
d7=Sat
d8=Sun
d9=Mon
d10=Tue
d11=Wed
d12=Thu
d13=Fri
d14=Sat
d15=Sun
Component 3
Component 2
Component 1


Loading Plot - 3D: Individual 10
Legend
d1=Fri
d2=Sun
d3=Mon
d4=Tue
d5=Wed
d6=Fri
d7=Sat
d8=Sun
d9=Mon
d10=Tue
d11=Wed
d12=Thu
d13=Fri
d14=Sat
d15=Sun
d16=Mon
Component 1
Component 2
Component 3


Loading Plot - 3D: Individual 15
Legend
d1=Tue
d2=Thu
d3=Fri
d4=Sat
d5=Thu
d6=Fri
d7=Sat
d8=Sun
d9=Mon
d10=Tue
d11=Wed
d12=Thu
d13=Fri
d14=Sat
Component 1
Component 2
Component 3


Loading Plot - 3D: Individual 17
Legend
d1=Tue
d2=Sat
d3=Mon
d4=Tue
d5=Thu
d6=Fri
d7=Sat
d8=Sun
d9=Mon
d10=Tue
d11=Wed
d12=Fri
d13=Sat
d14=Sun
d15=Mon
Component 1
Component 2
Component 3