Filtering and clustering GPS time series for lifespace analysis

by

Laura May Morrison
B.Sc., University of Victoria, 2010

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Filtering and clustering GPS time series for lifespace analysis

by

Laura May Morrison
B.Sc., University of Victoria, 2010

Supervisory Committee

Dr. Roderick Edwards, Co-supervisor
(Department of Mathematics and Statistics)

Dr. Julie Zhou, Co-supervisor
(Department of Mathematics and Statistics)

Supervisory Committee

Dr. Roderick Edwards, Co-supervisor
(Department of Mathematics and Statistics)

Dr. Julie Zhou, Co-supervisor
(Department of Mathematics and Statistics)

ABSTRACT

This thesis focuses on various aspects of community mobility and lifespace. Mobility is of particular interest to those working with the elderly population or patients affected by neurological diseases, such as Alzheimer's and Parkinson's diseases. One aspect of mobility is the number of "hotspots" in a person's daily (or weekly) trajectory, which represent the locations at which an individual remains for a minimum predetermined length of time. The individual demonstrates potential limited mobility if there is only one identified hotspot; the individual is more mobile if there are multiple identified hotspots. Based on GPS time series, we can use cluster analysis to identify hotspots. However, existing clustering algorithms such as k-means and trimmed k-means do not take into account the time dependencies between the location points in the series, and require knowing the number of clusters ahead of time. Thus, the resulting clusters do not represent the subjects' activity centres well. In this thesis we have developed a robust time-dependent clustering criterion that works very well to find clusters. Another aspect of mobility is the total distance travelled. The total distance computed from the original GPS data is inflated as there is noise in the data. Due to the particular characteristics of noise specific to GPS time series, we have investigated the identification of noisy segments of data as well as smoothing techniques. The average amplitude of acceleration is proposed as an appropriate method to identify the large noise that occurs in GPS data. A multi-level trimmed means smoother is proposed as an appropriate method to filter the identified large noise. Three methods were investigated to determine an ellipse that identifies the spatial area an individual purposely moves through in daily life. The classical and robust 95% ellipses contain 95% of the points, but do not necessarily capture the distinct shape of the data. The minimum spanning ellipse over the series with all

points in each identified cluster reduced to each cluster's central value captures the shape of the data very well and is proposed as the most appropriate lifespace ellipse. Results are obtained and presented for the subjects available in the mobility study for the total distance travelled and a meaningful lower bound, the number of hotspots, the proportion of time spent in the hotspots, as well as the area of the classical 95% ellipse, robust 95% ellipse and minimum spanning ellipse. In the processing of the data, other problems that had to be addressed include obtaining appropriate estimates for the missing values and translating the time series from degrees of longitude and latitude to metres in the Cartesian (x, y) plane.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Community mobility of humans, which will be referred to as mobility throughout this thesis, is of interest to many, particularly those in the medical and health fields. It is of special interest and importance to those working with the elderly or patients affected by various neurological diseases. Using precise Global Positioning System (GPS) location points of people for extended periods of time can help produce ideas of the mobility of various participant types, such as elderly and adults with diseases.

This thesis will present various measures and concepts that give insight into the level of mobility an individual has, such as the total daily distance travelled, the number of locations (clusters) visited in the individual's daily movements and how far from home these clusters are. Previously developed clustering algorithms will be reviewed and a new robust time-dependent algorithm will be presented to find clusters in an individual's daily movement patterns.

Section 1.1 presents the concepts of mobility and lifespace, which are the motivating factors in this thesis. Section 1.2 gives information on the GPS coordinate system used in the experiment from which the data will be analyzed. Section 1.3 gives a brief description of the data sets. Section 1.4 contains the research problems

which are investigated and Section 1.5 summarizes the significant contributions made in this thesis.

1.1 Mobility and lifespace

Community mobility is defined by Baker, Bodner and Allman (2003) to be the movements of an individual in terms of location. Hence mobility is essential for people to maintain a good quality of life, as it is required for numerous daily activities ranging from cleaning and cooking to grocery shopping, going to medical appointments and visiting friends and family in the community.

Mobility is of great interest to many people working in the health and medical fields and of particular interest to those working with the elderly population or individuals affected by various neurological diseases. This is due to the fact that these are the groups of individuals are at large risk for declines in their level of mobility.

Boissy et al. (2011) discuss how mobility studies used to be based on an individual's daily activity as reported by the individual and how these methods may not lead to the best results due to potential inaccuracies and bias in the reported movements. However, with the progression of GPS devices, researchers can now monitor the locations of individuals for an extended period of time to gain more accurate ideas of the mobility of individuals.

Mobility in modern day life has lead to vast changes in the structure of urban cities and in turn the lifestyle of those living in urban environments. Furthermore, in the past few centuries, there has been an increase in the availability to connect to different places, cities and countries around the world, leading to a far larger area for individuals to be interactive with. As presented in Mello and Marandola (2005), a concept to gain a better representation of an individual's actually mobility

has been proposed by the French demographer Daniel Corgeau and is known as an individual's lifespace. Boissy et al. (2011) define lifespace as the proportion of time spent travelling, or equivalently the proportion of time spent in the hotspots, along with the spatial area an individual occupies in his/her daily life. Mello and Marandola (2005) discuss the fact people are living more dynamic lives as a result of the lifespace of individuals living in urban environments increasing. The spatial aspect of lifespace is informative, as it allows us to see how much an individual is moving, which in turn gives an indication of how mobile he/she is and how he/she interacts with his/her environment.

1.2 Global Positioning System

Longitude and latitude measure geographic location. Most maps display lines of latitude and longitude, which typically run horizontally and vertically, respectively.

Latitude lines are parallel and equidistant from each other. Each degree of latitude is approximately 111 kilometres apart; there is a slight variation due to the fact that the earth is not perfectly spherical. Degrees of latitude are numbered from 0° to 90° north and south. Zero degrees of latitude is located at the equator, whereas 90° north is the North Pole and 90° south is the South Pole.

The vertical longitude lines converge at the two poles and are widest at the equator. Degrees of longitude are numbered from 0° to 180°, where 0° longitude is located at the Royal Greenwich Observatory in England. The point at which the zero degrees latitude and zero degrees longitude lines intersect is in the Gulf of Guinea in the Atlantic Ocean, which is approximately 611 kilometres south of Ghana and 1078 kilometres west of Gabon.

1.3 Data sets

The data that will be analyzed in this thesis were collected as part of a project on mobility and aging. The project involves monitoring activity by means of a wearable data logging platform that has GPS tracking abilities to assess the lifespace and mobility profile of individuals. The data sets are provided by a research team led by Dr. Patrick Boissy (University of Sherbrooke, Department of Surgery) and Dr. Christian Duval (University of Quebec at Montreal, Kinanthropology department).

There are a total of 35 subjects whose data will be used and analyzed in this thesis. This was the total number of readable data files available at the commencement of analysis for this thesis, although more subjects are or will be participating in the mobility study led by Dr. Boissy and Dr. Duval. The number of available days for each subject ranges from 3 days to 8 days for each time period, and there are two time periods for most subjects. Table 1.1 summarizes the number of recorded days in each time period for the 35 subjects. The NA entries represent the time periods where there is no recorded data for the particular subject.

Table 1.1: Days per time period for each participant

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Period 1	7	4	8	6	6	5	7	6	6	8	8	8
Period 2	6	5	8	7	6	3	8	6	8	NA	7	8
Subject	13	14	15	16	17	18	19	20	21	22	23	24
Period 1	4	6	6	6	5	5	5	6	7	6	7	7
Period 2	6	7	NA	7	5	6	6	6	7	7	7	7
Subject	25	26	27	28	29	30	31	32	33	34	35	
Period 1	5	7	6	7	8	8	6	8	8	6	4	
Period 2	6	8	6	6	8	8	7	8	8	7	7	

1.4 Research problems

Given time series data of the form (latitude, longitude, time), various characteristics of an individual's mobility and the area in which he/she conducts his/her daily life can be investigated. In this thesis, various methods are developed to analyze some of the aspects of individuals' lifespace based on time series geographical data collected in Sherbrooke, Quebec, Canada.

Data of this form may not have a constant sampling rate, which makes analysis of the data more challenging, since most time series analysis methodology is developed based on the assumption that the time points are equally spaced. Furthermore, since the data used in this thesis is based on GPS coordinates, it is possible that the signal is lost while an individual is in a particular location. To overcome these challenges, one may use linear interpolation on the data to create a complete time series with a constant sampling rate giving equally spaced time points. The advantages and disadvantages of linear interpolation applied to location time series are discussed in Section 2.1.

The number of stops an individual makes throughout a day, the amount of time spent at the locations at which the individual stopped, and the lifespace (area) the individual covered in a particular day or over a longer period are of interest to health researchers. An example would be that an individual has the following route on a particular day: home, school, work, school, bank, grocery store, home. Hence, the individual may stop at the same place more than once in a given day. If one counts each stop separately, home and school account for two stops each in the above example giving a total of 7 stops. However, if one is interested in the number of distinct locations the individual stopped, each location is accounted for only once giving a total of 5 stops for the previously described day. Robust methodology needs to be used in order to find where the centres of interest, also known as clusters or "hotspots",

are located and determine the time spent in these hotspots, as well as the distance each one is from the home location. These centres of interest are locations where an individual remains in approximately the same location for a minimum predetermined length of time. Robust methods are used rather than non-robust methods to get accurate estimates in the presence of possible outliers and noise, which are two very plausible situations with GPS data.

Another aspect of an individual's mobility that may be considered is the distance travelled throughout a given day or the proportion of time an individual was active or inactive. If this distance is small, the individual has not ventured out far beyond his/her home, whereas if this distance is large, the individual has likely either travelled a fair distance away from his/her home, or been steadily moving throughout the day. Similarly, if the proportion of time the individual spent in these clusters or hotspots is low, the individual has been fairly active that day, since much of his/her time has been spent on the move between the clusters.

Various methods will be explored in this thesis, including linear interpolation, k-means cluster analysis and trimmed k-means (robust) cluster analysis. All of these methods have been previously developed and applied to numerous data sets. We will apply these standard techniques for spatial clustering to our location data ignoring the time-dependencies among the data points to see how well these procedures work on our time-dependent location time series data.

In addition, methods which may be more suitable to time-dependent location data will be developed and investigated. We will present a clustering algorithm for time-dependent location data with the time variable included in the clustering. Analysis on the number of clusters or hotspots for the individuals will also be presented. Methods for computing the centres of the clusters and fitting ellipses around the individual clusters as well as the entire data set will be discussed as well. Various

methods of noise detection will be investigated and presented, as well as methods on how to smooth the data to obtain a less noisy data set and achieve more accurate results. Total distance estimates will be presented with lower bounds placed on the estimates.

1.5 Significant contributions

This thesis explores various aspects of mobility and lifespace of individuals in order to compare and classify mobility levels. Various complications of GPS location data are discussed and addressed. Properties of the hotspots are discussed and robust criteria for locating the hotspots are developed. Our algorithm is presented with output from several real life data sets collected in Quebec, Canada. Several methods for identifying large noise within the time series are explored. With knowledge of where the large noise occurs, various filtering techniques are investigated. Examples are presented to display the results from the large noise identification and filtering techniques. The methods presented in this thesis are limited to two-dimensional data, but may be extended to higher dimensions.

The thesis contains the following main contributions:

1. A new algorithm is developed to find clusters (hotspots) in time-dependent location data. This algorithm is very effective for the mobility data set.

2. Three methods of constructing ellipses are proposed to compute the total area covered by each individual, also known as their lifespace.

3. A new method is proposed to detect large noise in the data set and an effective filtering technique is developed.

4. Several measures are proposed to compute the total distance travelled including

a lower bound of the total distance. The lower bound and the total distance provide additional information about the noise level of each time series.

5. Various results of lifespace are presented for the mobility study data (distance travelled, area covered, number of hotspots).

Chapter 2

Description of Data

The data analyzed in this thesis are time series data consisting of GPS location points for 35 subjects from the mobility project led by Dr. Patrick Boissy and Dr. Christian Duval previously mentioned in Chapter 1. The length of recordings were not the same for each participant in the study or even each day for the same individual. Most participants wore the wearable GPS device for two time periods, which ranged from three to eight days in length and were not necessarily the same length for the same participant. Within each time period, the days are roughly consecutive. The two separate time periods are not consecutive. Therefore, there are two time periods of recordings rather than approximately two weeks of consecutive recordings.

The data has a sampling rate of one second, meaning there should be GPS coordinates available for every second the participant had the GPS device on which was to be from when he/she woke up in the morning until he/she went to bed in the evening. The data analyzed was not necessarily split by date. Since we are interested in continuous time series, the days were split according to when the individual started recording that day and when they finished recording that day. There were many instances when an individual stayed up past midnight and therefore that day's

recording consisted of points past midnight and into a new date. The days were split when there was a gap of more than 2 hours between the latest recorded time point on one day and the earliest recorded time point the next day, or if the individual stayed up past midnight, the day was split the first time there was a gap of more than two hours that morning. Another occurrence was that some participants left their device on for twenty-four hours a day. When this occurs, the days are separated by date. Hence, an individual's day will be considered as having gone from midnight to midnight. This decision was made because we want to include the individual's minimal movement while at home in the evening as part of the same day as the activities they did earlier in the day. Since we do not know exactly when they went to sleep for the night and got up in the morning, breaking the days by date is reasonable for our particular research problem.

There are several issues that arose with the time series location data: (1) missing values, (2) multiple recordings and (3) noise and outliers.

There are missing values in the majority of the observed time series. Missing values can occur in these data sets for various reasons.

(a) One reason is the nature of the GPS devices, since they can lose their signal in certain areas, such as buildings and forested areas, and therefore no data will be recorded during those time periods.

(b) There is the possibility that the participant does not turn on his/her wearable device when he/she first gets up for the day and therefore some data points may be missing at the beginning of the series for that particular day. Similarly, the participant may turn off the device before the end of his/her day, making it such that data is missing after the last recorded location point. Since there is no way of knowing whether there are missing values before the first data point or after the last, this possibility is ignored.

(c) Another possibility is that the battery dies on the GPS device and therefore the signal is lost for several hours or possibly the rest of the day for that given participant on that particular day.

To deal with the missing values within the recorded time frame of the day, mainly for problem (a), linear interpolation is used to gain a full and evenly time spaced time series, as discussed in Section 2.1.

The second issue that arose with the data is multiple recordings for a given time point, which were not identical in location. When this situation arose, the first recorded location at those time points for the given series was used. This issue was extremely rare and when it did occur, it only affected approximately a minute of recordings.

The third issue that arose with the data was noise amongst the location points, and in particular large noise when the individual remained in roughly the same spot. By looking at the recorded series, it appears that when an individual remains in a location for a prolonged period of time, such as at home or in an office building, the observed location points had a tendency to drift or jump quite far away from all previously recorded location points and then return to the presumed accurate location of the individual. To deal with large noise and possible outliers, a new procedure is developed to identify time periods of large noise, and then a robust method is proposed to smooth the time series. The details are discussed in Chapter 6.

2.1 Missing data and interpolation

Suppose there is a set of location data of the form $(u_i, v_i, t_i),\ i = 1, 2, \ldots, n$, where u_i represents the latitude, v_i represents the longitude, and t_i is the time at observation i. Two situations may arise with such a data set: (a) $t_i - t_{i-1} =$ constant, and

(b) $t_i - t_{i-1} \neq$ constant for $i = 2, \ldots, n$. Most time series analysis and sampling techniques assume that observations are equally spaced in time. Thus, in the situation when $t_i - t_{i-1} \neq$ constant, one can convert the data set into a new one with equally spaced time points using methods such as linear interpolation. The concept of linear interpolation works well for movement location data when the number of time points being filled in is small, such as situation (a) on page 10. This is due to the fact that the individual must get from one location to the next. The simplest way to go between the two locations is by a straight line which is what linear interpolation does. However, linear interpolation must be used with caution with location data; if a large time period is missing, linear interpolation may not be the best due to the simple fact that an individual likely did not walk in a straight line. For instance, if the time period being interpolated is 60 minutes, the individual may have walked around a few city blocks; linear interpolation will not use this path but instead will show a straight line between the two points, which may not even be a possible route for the individual to have taken. By using linear interpolation, we may underestimate the distance travelled, but this at least gives a lower bound on the unknown true distance between the endpoints of the interpolation. Linear interpolation will be used throughout this thesis with the knowledge that it may not be representing the exact route of the individual. The new observations are denoted by (u_i^*, v_i^*, t_i^*), where t_i^* is the time point that is desired and $t_i^* - t_{i-1}^* =$ constant, u_i^* and v_i^* are the latitude and longitude values based on linear interpolation between the two closest points in time on either side of t_i^*, respectively.

From Arden and Astill (1970), the values u_i^* and v_i^* for any given $t_i^* \in (t_{i-1}, t_i)$ are computed as follows,

$$u_i^* = u_{i-1} + \frac{(t_i^* - t_{i-1}) \times (u_i - u_{i-1})}{t_i - t_{i-1}} \quad \text{and} \tag{2.1}$$

$$v_i^* = v_{i-1} + \frac{(t_i^* - t_{i-1}) \times (v_i - v_{i-1})}{t_i - t_{i-1}}, \tag{2.2}$$

for any $2 \leq i \leq m$ where m is the length of the interpolated time series.

Example 2.1. *Figure 2.1 displays a data set, which is not part of the mobility study data presented in the remainder of the thesis, representing the GPS coordinates of an individual tracked during a daily routine, and Figure 2.2 gives plots of the latitude vs. time and longitude vs. time for the same data set. These data points are equally spaced for the most part with* $t_i - t_{i-1} = 5$ *seconds. However there are a few time periods where there are no recorded data points as shown in Figure 2.2. After applying the linear interpolation algorithm, we plot the equally spaced data points in Figures 2.3 and 2.4 with* $t_i^* - t_{i-1}^* = 5$ *seconds and* $t_i^* - t_{i-1}^* = 30$ *seconds, respectively.*

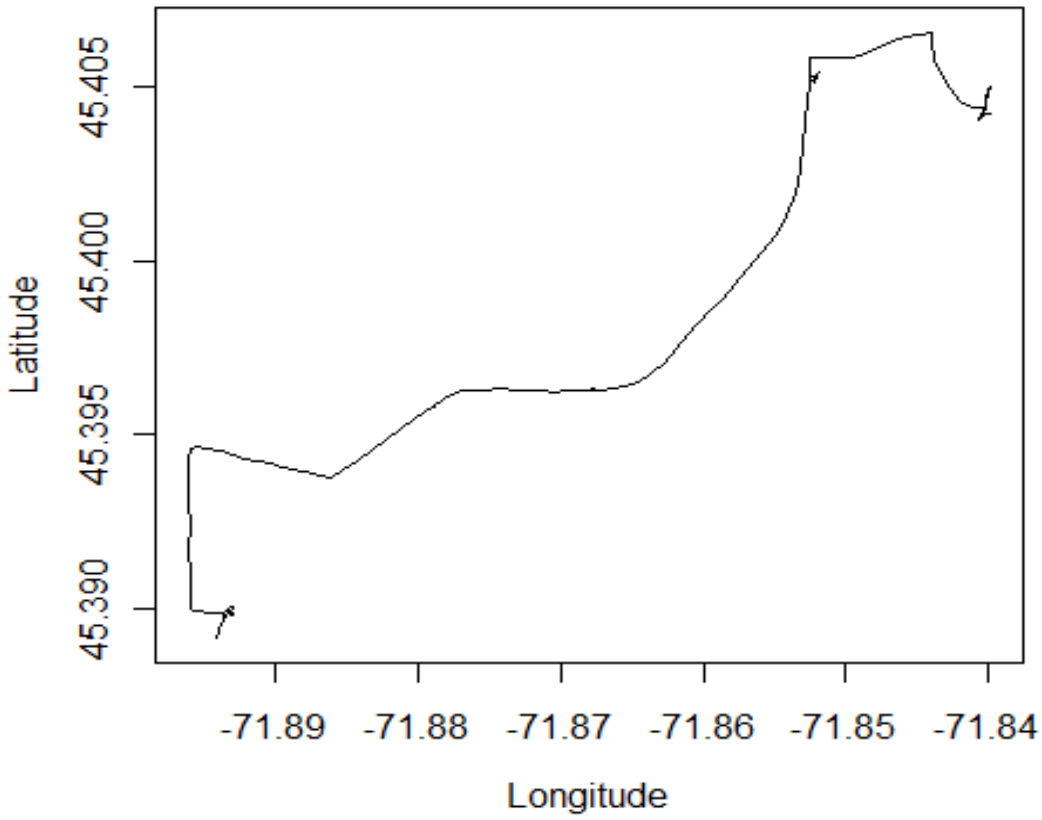

Figure 2.1: GPS coordinates of an individual's movements.

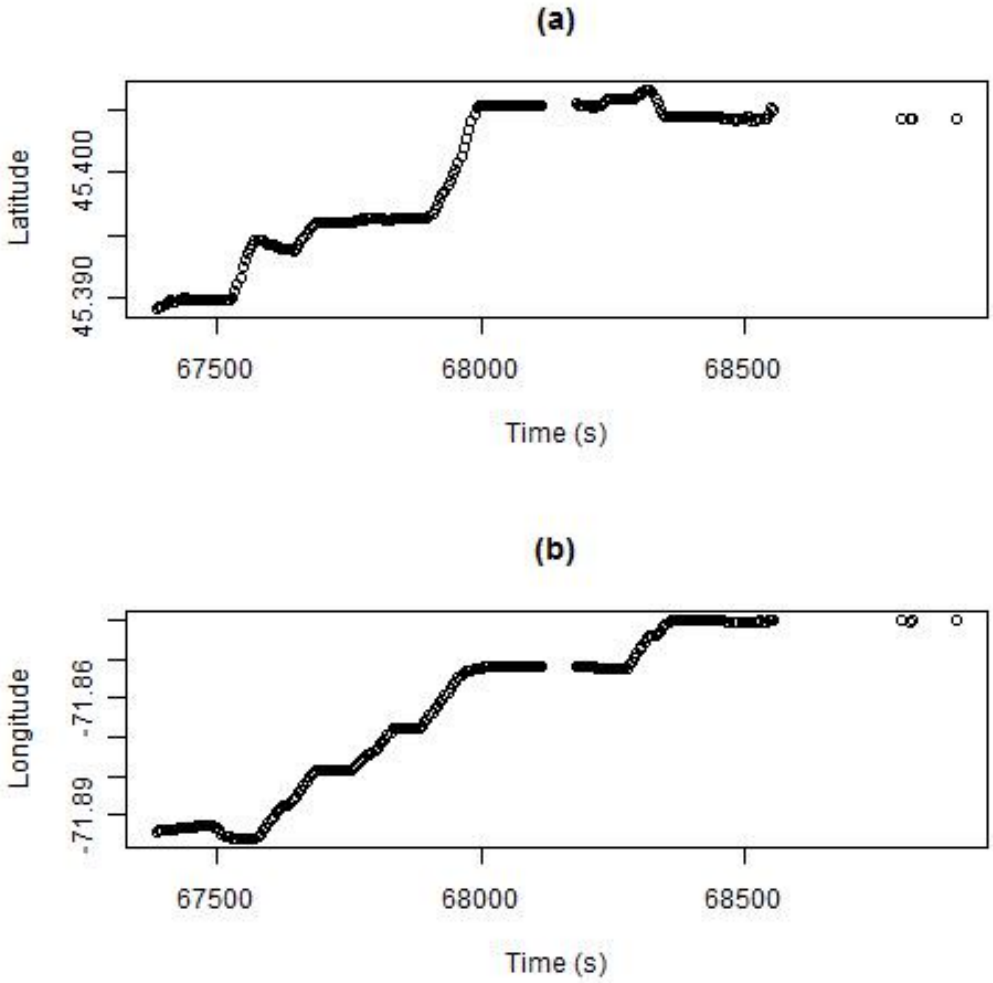

Figure 2.2: Movement plots: (a) Latitude versus Time, (b) Longitude versus Time.

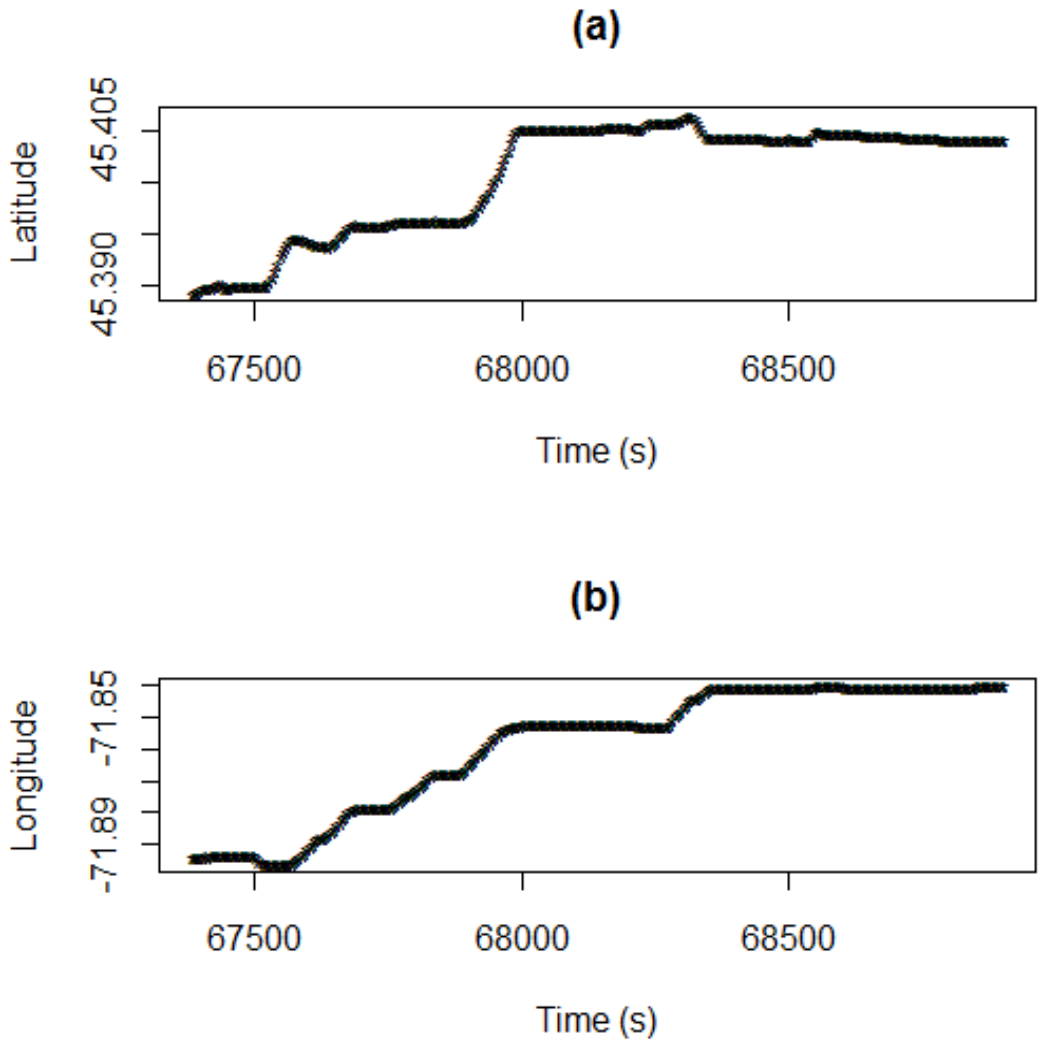

Figure 2.3: Movement plots using linear interpolation with $t_i^* - t_{i-1}^* = 5$ seconds: (a) Latitude versus Time, (b) Longitude versus Time.

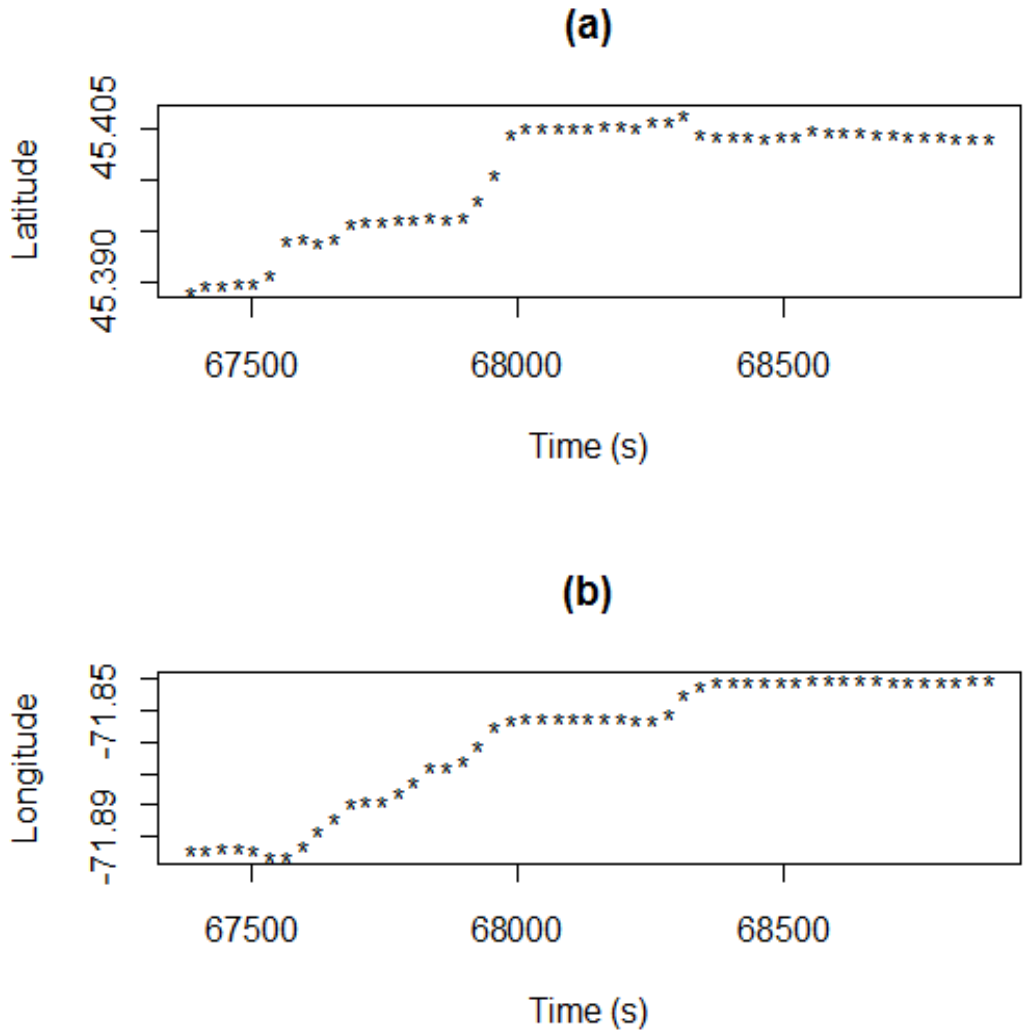

Figure 2.4: Movement plots using linear interpolation with $t_i^* - t_{i-1}^* = 30$ seconds: (a) Latitude versus Time, (b) Longitude versus Time.

2.2 Translate series from GPS coordinates to (x, y) Cartesian coordinates

In order to achieve a more easily understood measure of distance and area covered, all location points were converted into the (x, y) Cartesian coordinate system, where x and y are in metres, to carry out all further analysis. The error involved in converting from coordinates on a sphere to coordinates on a plane is minimal over the small distances that occur in these data, so the approximation is accurate enough.

This conversion was done in the following way. The first recorded data point for an individual on a given day is considered to be located at position $(x_1, y_1) = (0, 0)$, which is assumed to be the home location. From Rick (2004), the new value in the y direction for the i^{th} point is calculated as follows, $i > 1$,

$$y_i = 6371 \times 1000 \times (u_i - u_1) \times \pi/180 \text{ metres.} \tag{2.3}$$

Similarly, the new value in the x direction for the i^{th} point is calculated as follows:

$$x_i = 6371 \times 1000 \times (v_i - v_1) \times \pi/180 \times \cos(u_1 \times \pi/180) \text{ metres.} \tag{2.4}$$

Here, the value of 6371 represents the approximate radius of the earth in kilometres, the value of 1000 is to put the distance into metres and $\pi/180$ is used to convert the degrees into radians for the distance calculation. The x_i value always depends on the values of u_1 and v_1 due to the fact that we chose to calculate the distance between each point and the first location point rather than the distances between two adjacent points. This does not present a problem and is a reasonable approximation to the true value, as all the u_i and v_i values are extremely close to one another, respectively, in relation to Earth's curvature in a given time series for each individual.

Figure 2.5 depicts two location points, (u_1, v_1) and (u_2, v_2) and the geometry from which equations (2.4) and (2.3) are derived.

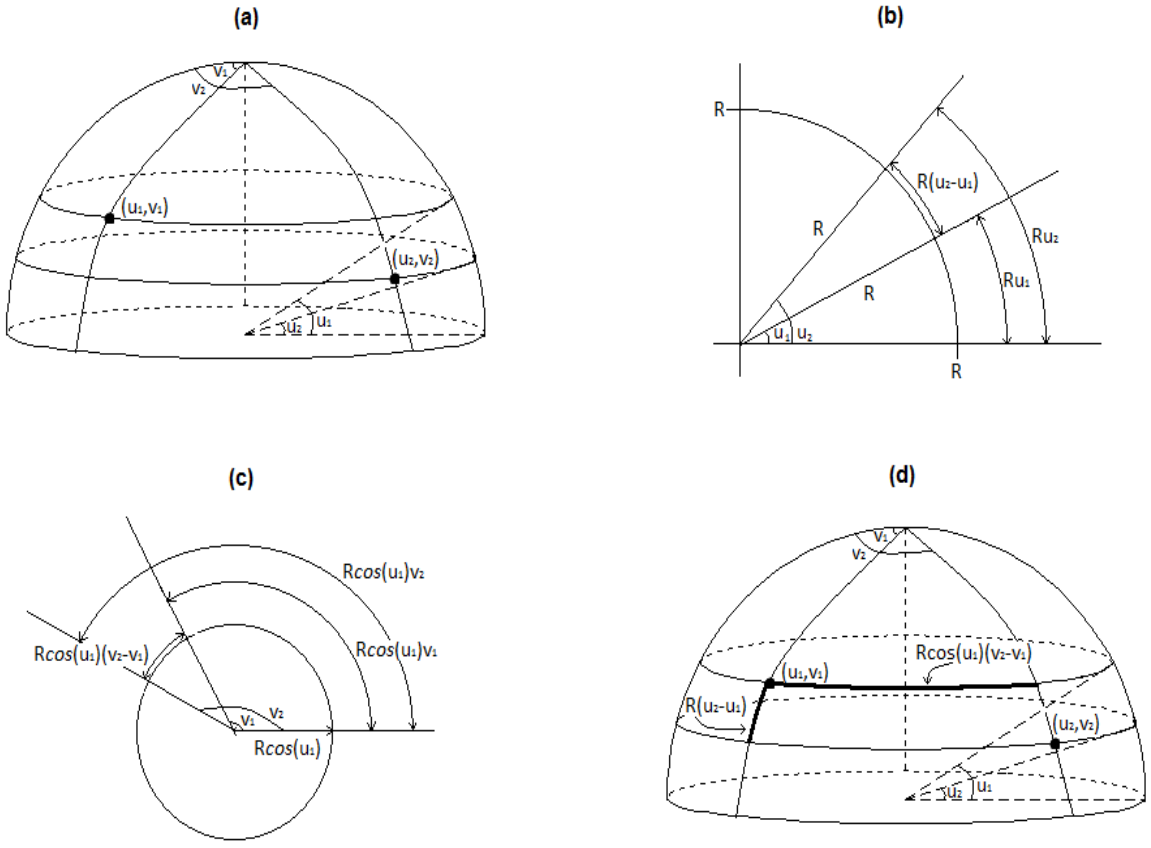

Figure 2.5: Diagram of how to get the equations of the translated series
(a) Graph of half of Earth with two identified location points, (u_1, v_1) and (u_2, v_2).
(b) Calculate the difference in latitude between (u_1, v_1) and (u_2, v_2) using arc length.
(c) Calculate the difference in longitude between (u_1, v_1) and (u_2, v_2) using arc length..
(d) Same graph as in (a) with the x and y distances between the two points in bold.

It can be seen that the distance between the lines of latitude remain constant. Furthermore, it can be shown that a one degree change in latitude is approximately 111 kilometres, i.e. $6371 \times 1 \times \pi/180 \approx 111.19$ kilometres.

Figure 2.6 displays the location points for subject 1, time period 1, day 2. The graph in Figure 2.6 (a) shows latitude vs. longitude of the interpolated series and the

graph in Figure 2.6 (b) shows y vs. x in the Cartesian coordinate system. It is clear that the shape has been preserved.

2.3 Summary statistics

To gain some perspective on our data, two summary statistics were computed: proportion of data recorded and total distance travelled.

The proportion of data points recorded will be considered as the ratio of the number of recorded data points to the number of possible points between the first and last data point in the series, where the number of possible points is equivalent to the number of seconds between t_1 and t_n. This proportion can be calculated as follows: $p_r = \#$ of recorded location points / $\#$ of possible points. Thus, $1 - p_r$ is the proportion of missing data. A low value of p_r indicates that there are many missing values and therefore inferences should be made with caution.

The total distance travelled by an individual throughout the day is calculated by $d_T = \sum_{k=1}^{m-1} \sqrt{(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2}$, where m is the number of points in the interpolated series. This allows us to get an approximate idea of how mobile an individual is without looking at his/her specific daily movements.

The proportion of recorded data, p_r, for all the time series can be found in Appendix A. This table also includes the total distance of the recorded trajectory, d_T, each participant travelled each day, as well as the length of time of each interpolated time series.

Figure 2.7 (a) displays the total distance travelled by subject 2 on each recorded day. The figure combines the two time periods and uses the measurements for all recorded days for this particular subject. The circular dots indicate exactly how far the individual went on a particular day with a vertical line going from zero to the

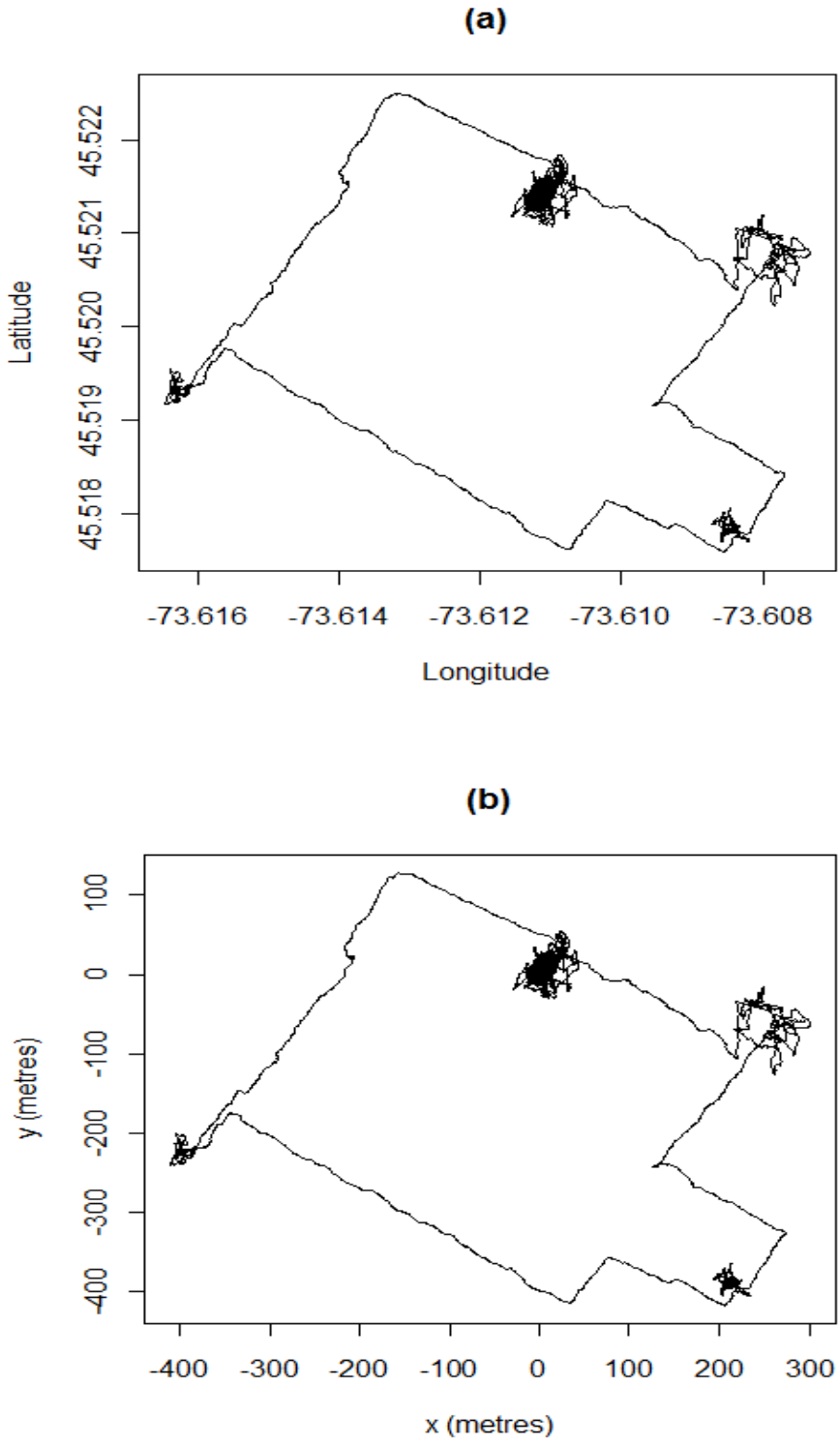

Figure 2.6: Location points for subject 1, time period 1, day 2:
(a) Latitude vs. Longitude, (b) y vs. x in Cartesian coordinate system.

total daily distance to gain a visualization of the distances travelled each day. Since the figure displays the distances for all recorded days for subject 2, the 9 available days of data are not consecutive due to the two time periods not being consecutive. The participant does not travel the same distance each day, which is clearly shown by Figure 2.7 (a) where the individual went roughly 200 km on one day and a little more than 0 km another day. This implies that the individual went quite far away from home some days and stayed home on other days. Figure 2.7 (b) displays the proportion of recorded data points, p_r, for each recorded day for subject 2. Since these are proportions, all values will be between zero and one. The proportions of recorded data for the available days of subject 2 vary from approximately 60% to 100%. A proportion closer to 1 is desirable, as it indicates that not many location points were interpolated to fill in the series. For instance, day 4 has very few interpolated points, whereas day 5 has nearly 40% of the series interpolated.

Figure 2.8 displays boxplots for the daily distances travelled and proportion of data points recorded for each individual in the study. Each "box" includes the daily distances/proportions for all the days in the two time periods of recordings for each individual. From Figure 2.8 (a), it can be seen that individuals have variability in the distances that they travel in any given day with some extreme values (approximately 0km to 600km for the various participants). A few of these large distance measurements may be the result of some extreme outliers in the recorded location points. From Figure 2.8 (b) it can been seen that some participants had the entire day recorded, whereas there were some days where much of the series was missing giving extremely low proportions of recorded location points for certain individuals. This is certainly not an ideal case, as it indicates there is much of the series where we have no information and have had to fill it with interpolated data points. From the boxplots it appears as if the median distance travelled and the median proportion of

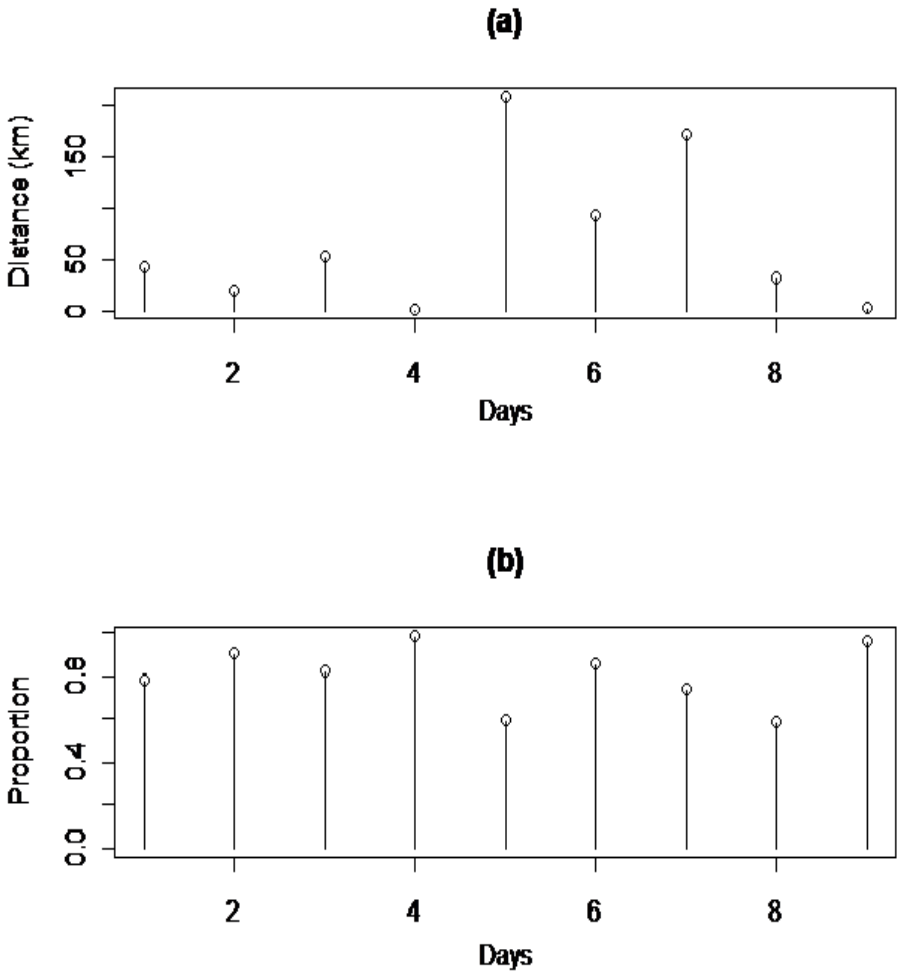

Figure 2.7: Measurements for subject 2:
(a) Total distance travelled (d_T), (b) Proportion of recorded points (p_r).

points recorded are not the same among all the participants in the study. Table 2.1 displays the median distances travelled and the median proportion of points recorded for each subject. As with the boxplots, it appears from the values in the table that the medians are not equal for either of these measures. To test this fact, the Kruskal-Wallis rank sum test was performed in the statistical software R using the function kruskal.test(). The test of equal medians for the distance travelled resulted in a Kruskal-Wallis chi-squared test statistic of 103.1505 on 34 degrees of freedom, giving a p-value of 6.852×10^{-9}. Therefore, the median distance travelled is not the same for all subjects. The χ^2 test was performed on the median proportions of recorded data points, which resulted in a test statistic of 2562.821 on 34 degrees of freedom and a p-value $< 2.2 \times 10^{-16}$, meaning that the median proportions are not all the same.

Table 2.1: Median distance travelled and median proportion of recorded points

Subject	1	2	3	4	5	6	7	8	9	10	11	12
d_T	27.7	43.4	14.0	36.0	46.6	58.6	20.4	25.9	10.1	24.0	58.6	45.9
p_r	0.72	0.82	0.58	0.67	0.72	0.74	0.47	0.80	0.67	0.54	0.58	0.50
Subject	13	14	15	16	17	18	19	20	21	22	23	24
d_T	15.5	20.3	25.6	42.4	14.3	24.5	28.1	44.0	21.5	28.9	25.8	41.1
p_r	0.81	0.70	0.46	0.84	0.89	0.81	0.87	0.73	0.49	0.73	0.72	0.68
Subject	25	26	27	28	29	30	31	32	33	34	35	
d_T	46.3	27.3	11.9	23.3	18.0	33.1	17.9	48.0	46.9	34.4	37.0	
p_r	0.93	0.76	0.62	0.54	0.64	0.45	0.75	0.73	0.68	0.81	0.65	

This chapter focused on dealing with missing data and converting the latitude and longitude location onto an (x, y) plane. Now we have a brief idea of the mobility of an individual through the distance travelled as represented by the total length of the recorded trajectory, and the quality of the data through the proportion of data available. Since there are large noise and outliers in the data, time series will be smoothed in Chapter 6 and the total distance travelled will be recalculated.

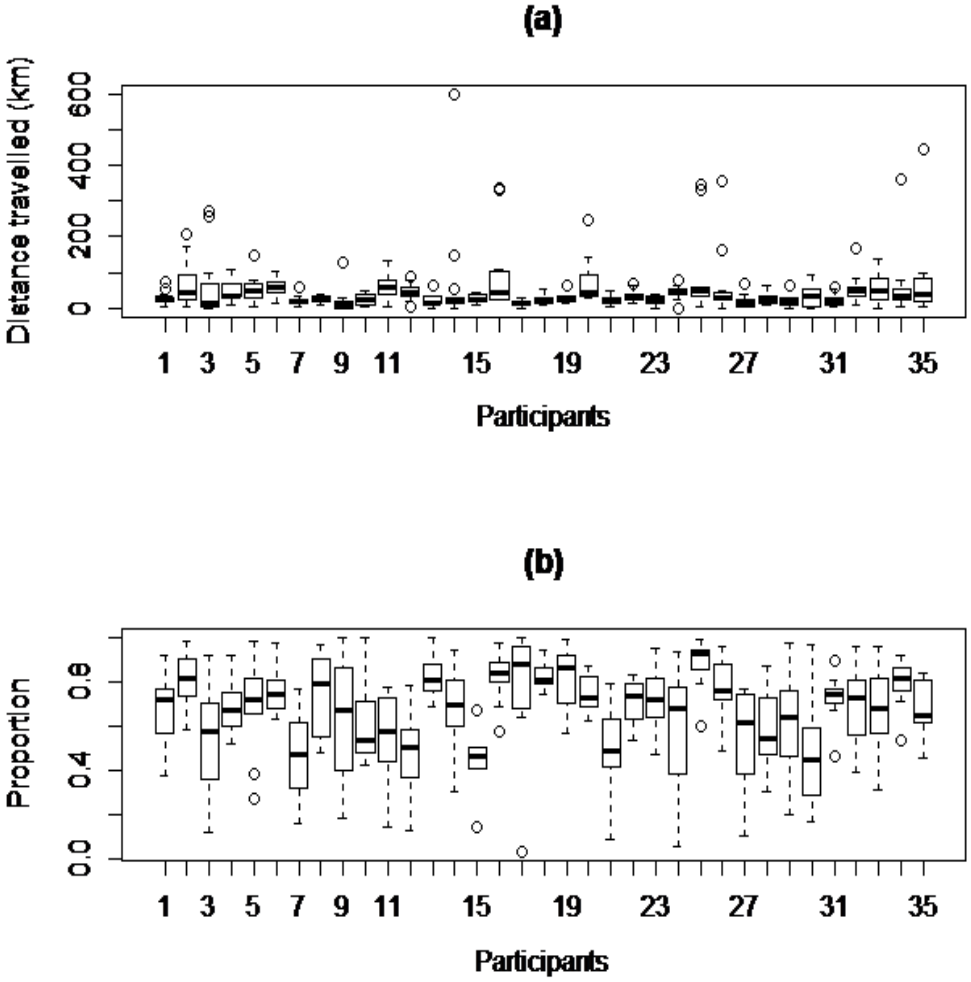

Figure 2.8: (a) Boxplot of distances travelled by all individuals,
(b) Boxplot of proportions of recorded data points for all individuals.

Chapter 3

Clustering Procedures

In this chapter, two standard clustering procedures for good data sets are reviewed. Section 3.1 gives the description of the k-means clustering technique. Section 3.2 discusses the trimmed k-means clustering procedure, which is a robust clustering technique. Several examples are given to find clusters.

3.1 k-means cluster analysis

Clustering techniques are designed to find k groupings of n data points, known as clusters, such that the units within groups are more "similar" than the units across groups. Gnanadesikan (1977) discusses how in most of these clustering procedures, the groups or clusters are determined by the iterative seeking of neighbourhoods that are defined in terms of some metric; that is, similar units are conceptualized as those that are close together in terms of some metric. In the case of our location data, we will use Euclidean distance as the metric. A popular non-hierarchical clustering technique is k-means clustering, which separates the data into k clusters, where $k < n$.

The Euclidean distance between two location points (x_i, y_i) and (x_j, y_j) is defined

by

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{3.1}$$

Suppose we have a data set (x_i, y_i), $i = 1, 2, \ldots, n$, that we would like to partition into k clusters, $C_1, C_2, \ldots, C_k$. As presented in Gnanadesikan (1977), the cluster centres, $(\bar{x}^s, \bar{y}^s)$, $s = 1, 2, \ldots, k$, are defined as

$$\bar{x}^s = \frac{1}{n_s} \sum_{i \in C_s} x_i \text{ and } \bar{y}^s = \frac{1}{n_s} \sum_{i \in C_s} y_i,$$

where n_s is the number of observations in cluster C_s. The Euclidean distance from an observation (x_i, y_i) to the cluster centre $(\bar{x}^s, \bar{y}^s)$ is denoted by

$$d_i(s) = \sqrt{(x_i - \bar{x}^s)^2 + (y_i - \bar{y}^s)^2}.$$

The k-means algorithm aims to divide n points into k distinct clusters such that the within-cluster sum of squares, $\sum_{s=1}^{k} \sum_{i \in C_s} d_i^2(s)$, is minimized. The algorithm used in this thesis was developed by Hartigan and Wong (1979), which seeks solutions such that no movements of a point from one cluster to another will reduce the within-cluster sum of squares.

Numerous suggestions have been given as to how to form the k starting points used as initial estimates of cluster centres. Dillon and Goldstein (1984) list a few of the proposed methods which include:

1. Choose the first k observations in the sample as the initial k cluster mean points.

2. Choose k observations that are mutually furthest apart.

3. Choose k initial cluster configurations based on prior knowledge.

In this thesis, the statistical software R will be used for data analysis. In this par-

ticular software, the k initial points for the cluster centres are k randomly selected points from the data set.

The k-means clustering method was nicely summarized by Afifi, Clark and May (2004) for a specified number of clusters k as follows.

1. Divide the data into k initial clusters. The members of these clusters may be specified by the user or may be selected by the software program.

2. For each of the k clusters, calculate the means or centroids.

3. For a given observation, calculate its distance to each centroid. If the observation is closest to the centroid of its own cluster, leave it in that cluster, otherwise, reassign it to the cluster whose centroid it is closest to.

4. Repeat step 3 for each observation.

5. Repeat steps 2, 3 and 4 until no observations are reassigned.

From this we can see that the value of k must be determined before the k-means procedure is used. Since the number of clusters is not always known, one must determine the best value for k.

Due to the performance of the k-means clustering algorithm being affected by the chosen value of k, it may be beneficial to use a set of values rather than a single value of k if there is no natural choice of k. The selected values must be significantly smaller than the number of points in the time series, which is the main motivation for performing data clustering. As discussed by Pham, Dimov and Nguyen (2005), the validity of the clustering result is often only addressed visually without applying formal performance measures. However, this approach can be difficult when clustering multi-dimensional data sets. Visual verification is largely applied due to it's simplicity and explanation possibilities.

In order to use this algorithm, a few assumptions are made. It is assumed that an appropriate value of k is used in determining the clusters. The k-means procedure also makes assumptions such as local independence or equal within-class variance. Furthermore, due to the fact that the squared Euclidean distance is being used, there is an implicit assumption that the data should have roughly the same scale to use such distances. Thus, in the case of location data, the x direction and y direction should be in the same metric. This means, for example, that if you have two data points, (x_1, y_1) and (x_2, y_2), you would not consider the change between x_1 and x_2 in kilometres and the change between y_1 and y_2 in metres, but rather both in kilometres or both in metres.

Example 3.1. *Consider a simulated data set with* $n = 326$ *shown in Figure 3.1. It can be seen that there are two clear clusters as well as observations forming a so-called path between the two clusters, which in some circumstances may be considered as its own cluster of observations. Statistical software R function kmeans() is applied to find the clusters. Figure 3.1 displays the data set after applying the k-means procedure using* $k = 2$ *and* $k = 3$*. It can be seen that when 2 clusters were used, the two distinct clusters are identified in separate clusters, but both are also clustered with part of the so-called path. When three clusters were used, the two distinct groups of observations are identified as clusters, and the majority of the trail is identified as a third and separate cluster.*

Example 3.2. *Consider the following real-life example of location data in Figure 3.2 obtained from Dr. Patrick Boissy's mobility study. These are the location points for subject 1, time period 1, day 2. This time series has* $n = 13546$ *observations for* 17637 *seconds. From the plot, it appears that there are four locations where the observations are grouped together and points joining the four groups into one continuous time series. The k-means procedure did fairly well at separating the large groups of points*

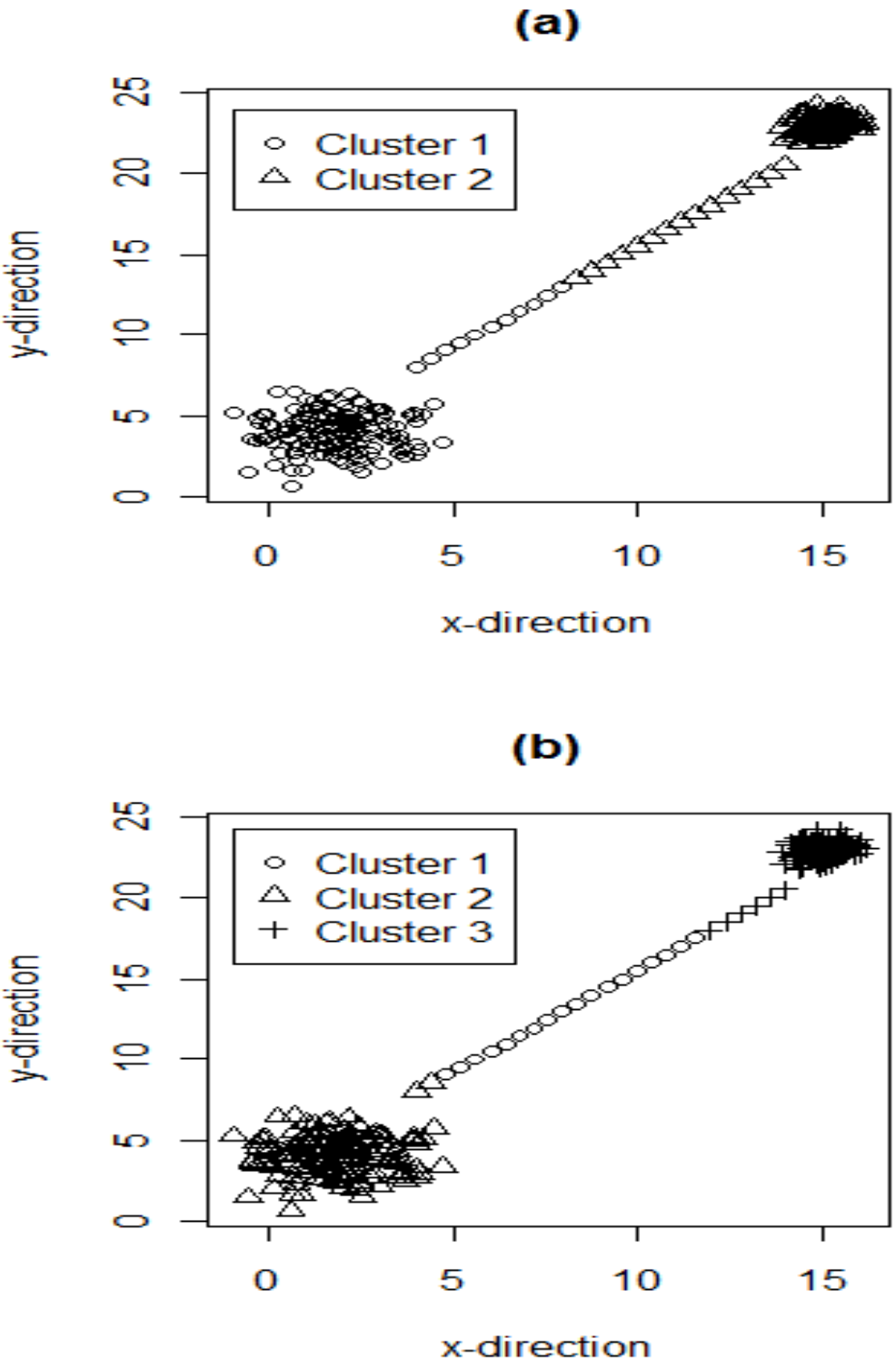

Figure 3.1: Plots of data separated into clusters using k-means analysis: (a) $k = 2$, (b) $k = 3$.

from each other. However, the points that join the groups together have been assigned to clusters though they look as if they are more of a joining path than part of a cluster. This may not be the most desired result for such location data, as one topic of interest is how long an individual is located at the clusters and therefore points that join the clusters should not be included.

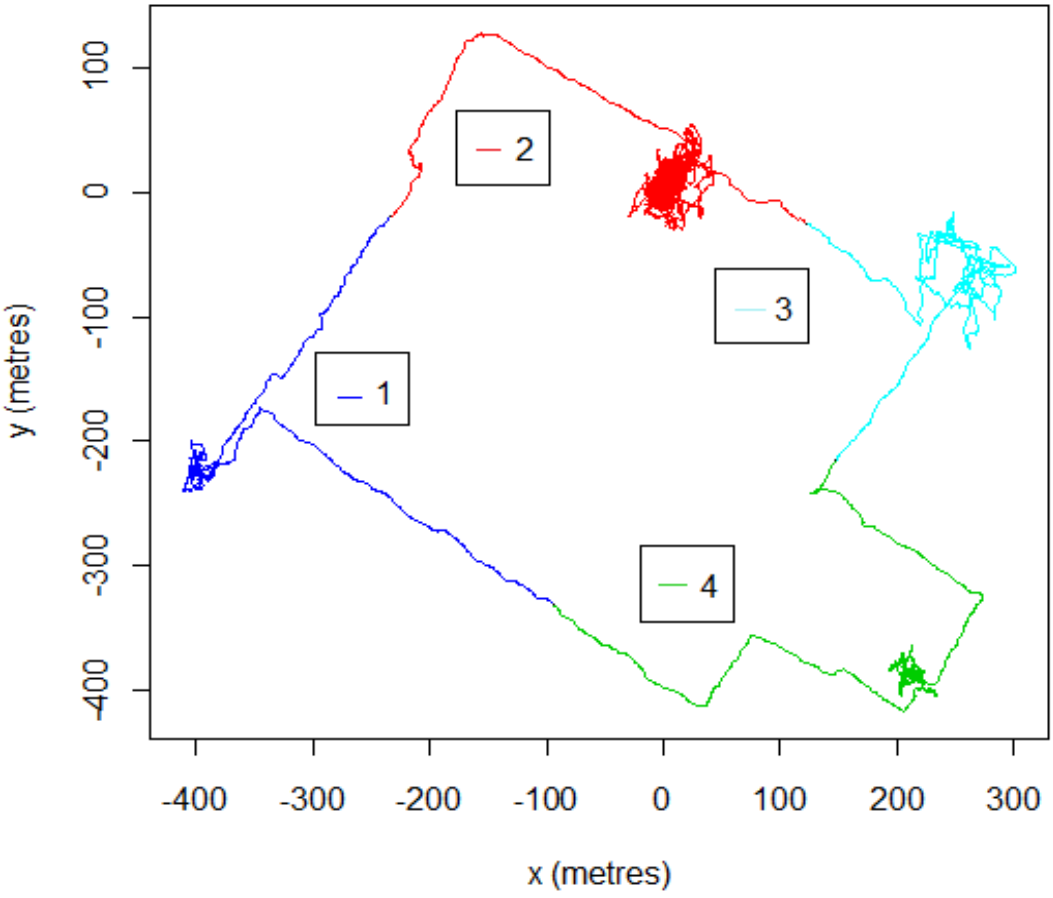

Figure 3.2: k-means clustering of subject 1, time period 1, day 2 location points.

3.2 Robust k-means cluster analysis

In Section 3.1 the k-means procedure was applied to a few examples of two-dimensional data. It was noted that some observations were assigned to a certain cluster when

the observations appear to not belong in a particular cluster but rather a trail between two clusters or a potential outlier for the cluster in question. To help deal with these undesirable situations that may arise in the classical k-means procedure, we will consider a robust version of the k-means procedure.

Generalized trimmed k-means was first introduced by Cuesta-Albertos, Gordaliza and Matran (1997). The idea behind it is that one must not give the same importance to a natural data cluster as an artificial cluster attributable to the presence of a small proportion of outliers.

Given a trimming level $\alpha \in (0, 1)$, define $n_\alpha = \lfloor n(1-\alpha) \rfloor$, where $\lfloor z \rfloor$ denotes the greatest integer that is less than or equal to z. A generalized trimmed k-means produces k clusters, $C_1, C_2, \ldots, C_k$, such that the following function is minimized:

$$\min_{Y} \sum_{s=1}^{k} \sum_{\substack{(x_i,y_i)\in C_s \\ (x_i,y_i)\in Y}} \Phi(d_i(s)), \tag{3.2}$$

where the set Y includes n_α points from the data set $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. From Garcia-Escudero and Gordaliza (1999), it can be seen that this is the k-mean of the subsample containing $\lfloor n(1-\alpha) \rfloor$ points with the smallest mean deviation penalized through the function Φ. The penalty function, $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$, is assumed to be continuous, strictly increasing and such that $\Phi(0) = 0$ and $\Phi(x) < \Phi(\infty)$ for all x. The penalty function $\Phi(x) = x^2$ is used in this thesis. In the statistical software R, there is a function trimkmeans() to compute the generalized trimmed k means.

Example 3.3. *Consider the real-life example of location data discussed previously in Example 3.2. It was shown that when the data is classified into four clusters using the k-means clustering technique, each of the clusters includes observations that are clearly not part of a true natural cluster. Figure 3.3 displays the result of clustering the data set using the trimmed k-means procedure as discussed above with trimming levels*

$\alpha = 10\%$ and $\alpha = 20\%$. When the trimming level $\alpha = 10\%$ was used, four distinct clusters were found in the four positions. These clusters no longer include most of the data points that join the clusters, which is far more desirable for our location data than the result from the k-means procedure. However, it can be seen that there are still a few undesirable properties presented in the plots. In Figure 3.3 (a), some of the points that look as if they should belong to a cluster have been trimmed, which can be seen with points located around Clusters 3. On the other hand, Cluster 4 seems to contain points that appear to be part of a path joining clusters 1 and 4, as well as clusters 3 and 4, and probably should not be classified as part of any cluster. When a trimming level of $\alpha = 20\%$ was used, the previously defined Cluster 3 is eliminated and the previously defined Cluster 2 gets separated into 2 clusters. This is clearly undesirable since a large grouping of location points is no longer identified and another grouping of location points has now been split into two clusters. Therefore, it is very critical to choose an appropriate trimming level α. In addition, it is also important to choose the number of clusters k that is appropriate and viable for the particular time series.

In this chapter, two clustering techniques were reviewed and examples were given. It seems as if the k-means clustering procedure is not very effective in identifying clusters/hotspots for time-dependent location data since it ignores the time component and includes all the points in the series rather than just those in hotspots. The trimmed k-means clustering technique did much better at identifying the clusters formed in the time series. However, it too ignores the time dependencies between the data points in the series. Moreover, it is an issue to choose k and α in the procedures. In the next chapter, a new clustering algorithm is developed that takes the time-dependent nature of the location time series data into account and identifies the clusters formed by an individual's movements throughout the day without the need to specify the number of clusters beforehand.

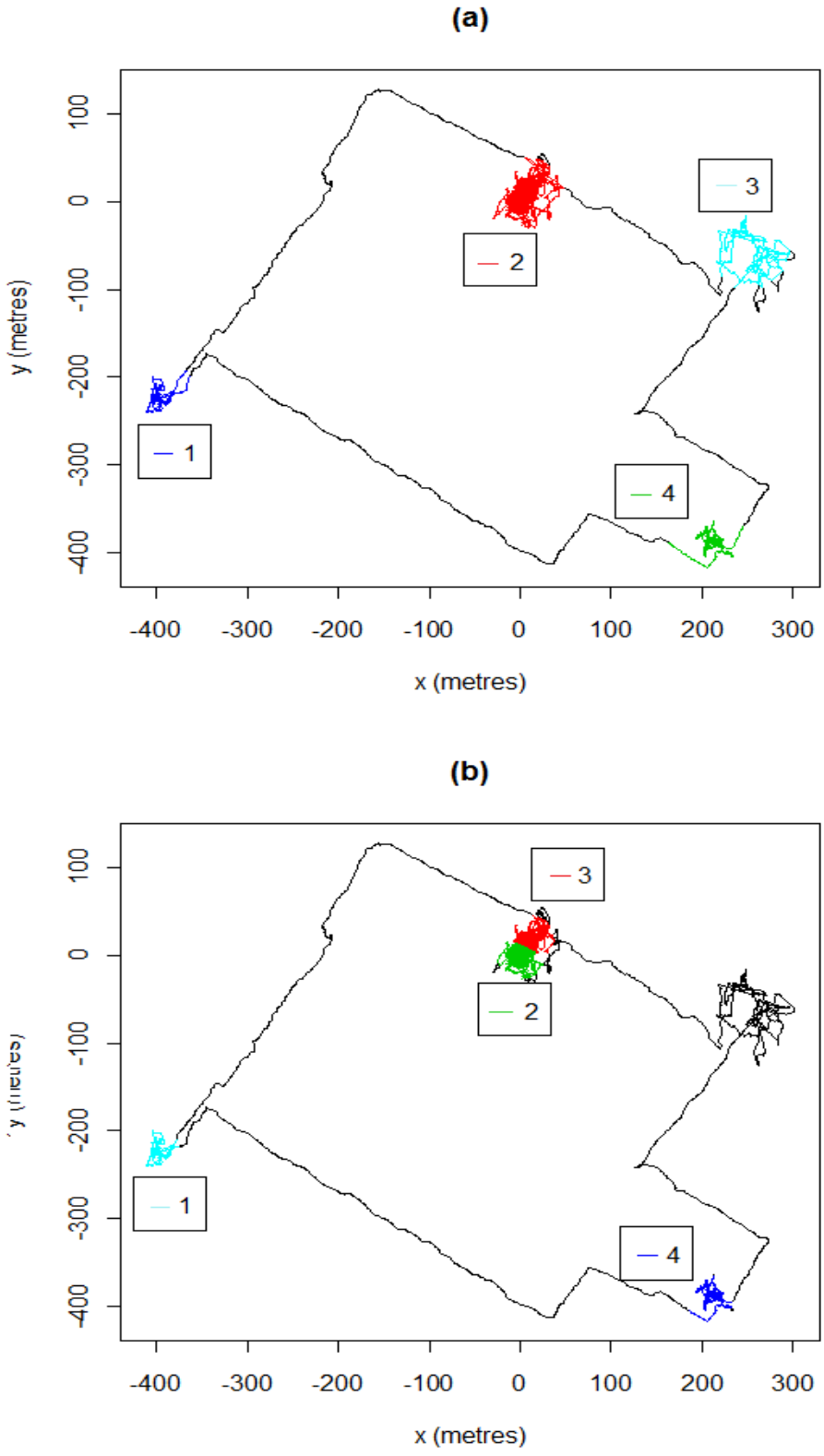

Figure 3.3: Trimmed k-means clustering for subject 1, time period 1, day 2 with: (a) $\alpha = 10\%$, (b) $\alpha = 20\%$.

Chapter 4

A New Clustering Procedure

As presented in previous chapters, one of the main objectives of this thesis is to develop a method to effectively determine where "hotspots" or clusters are located in a given individual's movements for a particular day. In this chapter a new method for identifying where these hotspots are located is proposed. The main idea for this new method is to have a time window scroll through the time series in fixed blocks of time looking for points that are located near each other in time.

4.1 Concepts and notation

We will now introduce several concepts and notation to describe a new clustering procedure for a time series of observations (x_i, y_i, t_i), $i = 1, 2, \ldots, n$. This method uses the idea of having a time window scroll through the time series and classifying whether each window represents points that are within a cluster.

Let $t_d = t_{j+1} - t_j$. Assume that t_d is constant. If t_d is not constant, the interpolation procedure in Section 2.1 is applied to get observations equally spaced in time. The i^{th} window will be denoted by w_i and the size of the time window is considered to be the length of time the windows span and is denoted by s. Therefore, if the time

windows are taken to be 5 minutes in length, then $s = 300$. In this new procedure, the windows are allowed to overlap one another by a ratio of $r \geq 0$.

To clarify this concept of the scrolling window, Figure 4.1 displays how the windowing would work if one was to use 5 minute windows ($s = 300$) that overlap the adjacent windows by half ($r = 0.5$). Thus, using $t_d = 1$, the first window, w_1, includes points 1 through 300, w_2 includes points 151 to 450, w_3 includes points 301 to 600, etc.

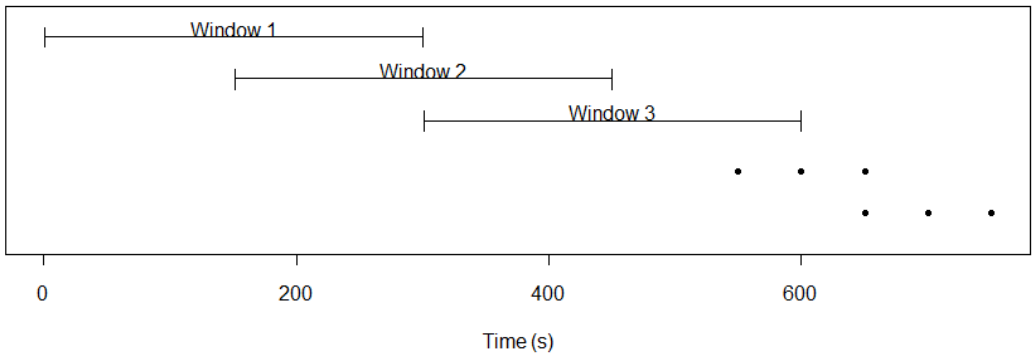

Figure 4.1: Five minute time windows with an overlap ratio of 1/2.

For a given window w_i, denote the centre of the window by (x_{c_i}, y_{c_i}), where x_{c_i} and y_{c_i} are the medians of the x and y values of the points in the window, respectively. To calculate this centre point for window i, let $a = s/t_d * (1 - r) * (i - 1) + 1$ and $b = s/t_d * (1 - r) * (i - 1) + s/t_d$. Then the centre location for window i is given by

$$x_{c_i} = median(x_a, x_{a+1}, \ldots, x_b) \text{ and } y_{c_i} = median(y_a, y_{a+1}, \ldots, y_b).$$

Various measures can be considered when trying to determine closeness of points in a given window. These may include the maximum distance from a centre point, the total distance travelled, the area covered or the $(1 - \gamma)^{th}$ quantile distance from each point to the centre point (x_{c_i}, y_{c_i}) for each window w_i.

The distance from each location point (x_j, y_j) in w_i to the centre point of window i is computed as follows:

$$d_{t_j} = \sqrt{(x_j - x_{c_i})^2 + (y_j - y_{c_i})^2}.$$

Since it is desirable to have a robust method of identifying whether a window is a cluster or not, we will use the $(1-\gamma)^{th}$ quantile of these distances d_{t_j}, $j \in w_i$. The quantile of these distances d_{t_j} will be denoted by $q_{1-\gamma,i}$ for window w_i.

To determine if window w_i is a cluster or not, compare the value of $q_{1-\gamma,i}$ to a predetermined cut-off value, say R. If $q_{1-\gamma,i} < R$, window w_i is considered a cluster of data points. Otherwise, window w_i is not a cluster.

Essentially, we are looking at a window of time, w_i, and forming a circle of radius R around the centre point (x_{c_i}, y_{c_i}). If the $(1-\gamma)^{th}$ quantile of the distances between all the points in window i and (x_{c_i}, y_{c_i}), $q_{1-\gamma,i}$, is within this circle, consider all the points in the window to form part of a cluster. If this value is outside of this circle centred around (x_{c_i}, y_{c_i}), consider the points in the window not to be from a cluster.

If adjacent windows have been identified as clusters, combine those windows into one larger cluster. For example, if there are 10 windows of time that have been scanned through, and windows 1, 5, 6, 7 and 8 have been identified as clusters, we will consider this data set to have 2 clusters, assuming the two clusters are in different locations. This is due to the fact that we have a cluster for window 1 and a cluster for windows 5 through 8.

4.2 Number of clusters and distance from home

The number of unique clusters is of interest in this thesis rather than the number of stops an individual makes. In order to determine the number of unique clusters,

the distances between the centre points of each of the clusters are computed. If the distances are very small or the points in the clusters overlap, those identified clusters are joined and labeled as one unique cluster.

Therefore, the overall algorithm is as follows:

1. Scroll through the time series with a moving window of length s, where the first window starts at the first data point in the series and the windows have an overlap ratio of r. In some examples, a window size of 5 minutes and overlap ratio of $r = 0.5$ are used.

2. For each time window w_i, find the median of the x values and the median of the y values of all the points in the window, denoted by (x_{c_i}, y_{c_i}). Notice that this may not be a location point in the time series.

3. Compute the Euclidean distance d_{t_j} from each point (x_j, y_j) in window w_i to the centre point (x_{c_i}, y_{c_i}).

4. For a given γ level, if $q_{1-\gamma,i}$ is less than the given acceptable radius, R, flag the window as a cluster of points. In some examples, $\gamma = 0.2$ and $R = 30$ metres are used. These were determined experimentally as appropriate values to use based on the location data. If $\gamma > 0$, the algorithm is robust against noise or outliers. The larger γ is, the more outliers the algorithm can deal with. There is potential for slight inaccuracy in this step as the clusters may not start or end precisely where the windows are starting or ending and therefore the identified window may include a few points that are the individual's location points as he/she travels between clusters.

5. Shift the window in time such that they overlap by the given ratio, r.

6. Repeat steps 2-5 until the end of the series has been reached.

7. If consecutive windows are flagged as clustered points, join them and classify the points as one larger cluster.

8. Compute the Euclidean distances between the centre points of each identified clusters. If they are within some predetermined distance of each other, say R_c, label those clusters as one unique hotspot location.

Health researchers may be interested in how far the hotspots are from the subject's home. This may tell the researcher something about the overall mobility capacity the individual has.

In this thesis, the distances between each hotspot and the home is considered to be the Euclidean distance between the centre of each identified unique cluster and the first data point in the series, $(0, 0)$, which is assumed to be the individual's home location. Therefore the distance between cluster i and the subject's home can be computed as

$$d_{h_i} = \sqrt{(x_{c_i} - 0)^2 + (y_{c_i} - 0)^2}. \tag{4.1}$$

If these distances are small, it implies the individual has not gone far from his/her home, which may imply the subject has limited mobility. On the other hand, if some of the distances between the hotspots and home are large, it demonstrates the subject has the mobility capacity to go great distances from his/her home.

4.3 Length of time spent in hotspots

Now that the number of hotspots for a time series of location data has been found, the length of time an individual spent in a given hotspot is of interest. For instance, an individual is likely to spend a great deal of time in the hotspots identified for locations such as the home or workplace, and not as much time for those identified

for locations such as a coffee shop, grocery store, shopping mall, etc.

The amount of time spent in cluster i may be calculated by $T_i = n_i \times t_d$, $i = 1, 2, \ldots, k$, where n_i is the number of data points in cluster i and t_d is the time difference between two consecutive time points. The total time spent in the clusters is calculated by $T_c = \sum_{i=1}^{k} T_i$ and the proportion of time spent in all of the clusters is calculated as $P_c = T_c/T$, where T is the total time of the time series.

For example, consider an individual has a continuous time series of location points for $n = 18000$ seconds (5 hours) with $t_d = 2$, for which 2 clusters were identified. Suppose the number of points in cluster 1 is determined to be 4500 and in cluster 2 is determined to be 1500. Then $T_1 = 4500(2) = 9000$ seconds (2.5 hours), $T_2 = 1500(2) = 3000$ seconds (50 minutes), $T_c = T_1 + T_2 = 12000$ seconds (3 hours 20 minutes) and $P_c = 12000/18000 = 2/3$.

4.4 Clustering results

Example 4.1. *Consider the data set in Examples 3.2 and 3.3. Figure 4.2 represents subject 1's particular movement for this given day. This figure displays a very distinct "city-block" appearance, which may be expected from an individual's daily movement patterns in an urban environment. Furthermore, it can be seen that there are four distinct areas which have been identified as clusters. This may or may not mean the individual only made four stops because the individual may have stopped at any of these identified clusters more than once. Since we are looking for the number of unique hotspots for this individual on this particular day, we would expect our algorithm to produce an answer of 4 hotspots based on the location points presented in Figure 4.2. In the data set, there are* $n = 13548$ *observations for 17637 seconds and* $t_d = 1$*. In the algorithm, we set* $s = 300$ *seconds,* $r = 0.5$*,* $R = 30$*, and* $\gamma = 0.2$*, which results*

in 4 clusters being identified.

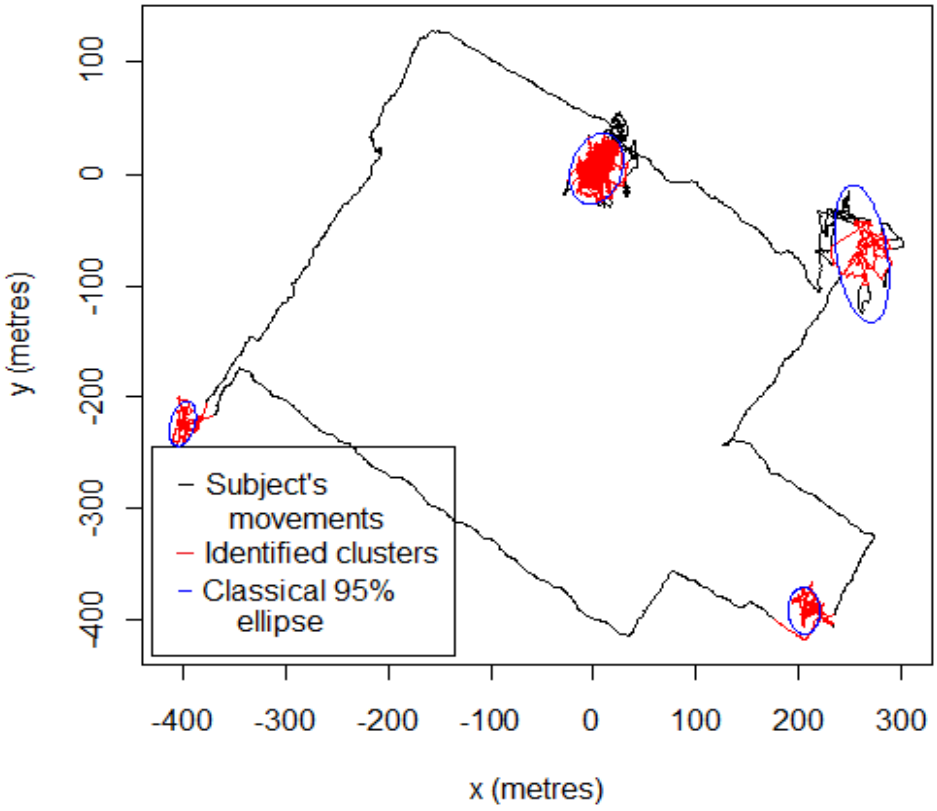

Figure 4.2: Plot of time series for subject 1, time period 1, day 2 clustered using the new robust time-dependent scrolling window method.

This data set for subject 1, time period 1, day 2 has now been clustered by the k-means procedure, trimmed k-means procedure and our new scrolling window clustering procedure. Table 4.1 gives some summary statistics for identified clusters. It is clear that the k-means procedure does not perform well since it does not allow us to discard the points that form the trails between the clusters and only look at the time points when the individual was in one location for an extended period of time. The trimmed k-means procedure outperformed the k-means procedure. It was able to identify four unique clusters that match what one would expect based on looking at the location points displayed in Figure 4.2. There were a few points that

were discarded when they perhaps should not have been and a few points that were included when they maybe should not have been. The new clustering algorithm was run with various values for the parameters to test whether the algorithm is sensitive to the choice of the parameter values. For most of the choices of parameter values, the new clustering algorithm was able to identify the four clusters. When $\gamma = 0.1$, $s = 300$ and $R = 30$ is used, one of the clusters is no longer identified. This is due to less of the further points being trimmed out in the comparison of the distance from each point to the centre point in the window. However, all the other combinations of the parameter values ($\gamma = 0.2$, $s = 300$, $R = 30$; $\gamma = 0.3$, $s = 300$, $R = 30$; $\gamma = 0.2$, $s = 300$, $R = 50$; $\gamma = 0.2$, $s = 420$, $R = 50$; $\gamma = 0.1$, $s = 180$, $R = 30$) have identified the correct location for the four clusters and the centre points are all very similar, along with the number of points in the clusters and the distance from the centre point in the cluster to the $(0, 0)$ location. Figure 4.3 displays the results from the six parameter combinations in the new clustering procedure. It can be seen that the algorithm is identifying the same locations and is not overly sensitive to the parameter value choices, apart from Figure 4.3 (f) in which one of the clusters was lost.

The cluster centres, (x_c, y_c), were fairly close for all three methods and parameter choices, as well as the distance from the centre of each cluster to the $(0, 0)$ location point. However, it is clear that the number of points in the clusters were very different between the k-means method in comparison to the trimmed k-means method and the new clustering method.

This new scrolling time window clustering procedure has some advantages over the other clustering procedures presented. The new method does not require k to be specified and is also robust against outliers. Furthermore, it is not overly sensitive to the choices of parameter values, meaning the results are similar and accurate for

reasonable choices of γ, s and R. In the next chapter, the new clustering technique is applied to all the data sets in the mobility study being analyzed in this thesis and results are presented.

Table 4.1: Summary Statistics for subject 1, time period 1, day 2

	Cluster	x_c	y_c	# Points in Cluster	Distance between cluster and (0,0) (metres)
New method	1	-399.7	-244.5	1566	468.6
$\gamma = 0.2$	2	3.7	4.8	7743	6.1
$s = 300$	3	261.1	-71.2	533	270.6
$R = 30$	4	204.9	-391.6	4548	442.0
New method	1	-398.7	-223.8	1568	457.2
$\gamma = 0.3$	2	4.1	5.2	7777	6.6
$s = 300$	3	258.3	-56.5	653	264.4
$R = 30$	4	204.9	-391.6	4548	442.0
New method	1	-398.7	-224.5	1586	457.6
$\gamma = 0.2$	2	3.9	4.9	8470	6.3
$s = 300$	3	258.3	-64.3	959	266.2
$R = 50$	4	204.9	-391.4	4594	441.8
New method	1	-400.0	-223.8	1612	458.4
$\gamma = 0.2$	2	4.0	5.6	8726	6.9
$s = 420$	3	259.5	-62.3	1082	266.9
$R = 50$	4	204.9	-391.5	4613	441.9
New method	1	-399.0	-224.7	1567	457.9
$\gamma = 0.1$	2	3.9	4.7	7749	6.1
$s = 180$	3	257.5	-54.7	687	263.2
$R = 30$	4	204.9	-391.5	4549	441.9
New method	1	-399.7	-224.5	1566	458.4
$\gamma = 0.1$	2	3.7	4.8	7743	6.1
$s = 300$, $R = 30$	3	204.9	-391.6	4548	442.0
Trimmed k-means	1	-398.3	-224.7	1598	457.3
$\alpha = 0.1$	2	4.4	5.1	8646	6.7
$k = 4$	3	253.5	-52.3	1028	258.8
	4	204.9	-392.0	4602	442.3
k-means	1	-396.8	-223.7	2059	455.5
$k = 4$	2	4.1	5.9	9205	7.2
	3	248.5	-57.9	1294	255.2
	4	204.9	-391.0	5079	441.4

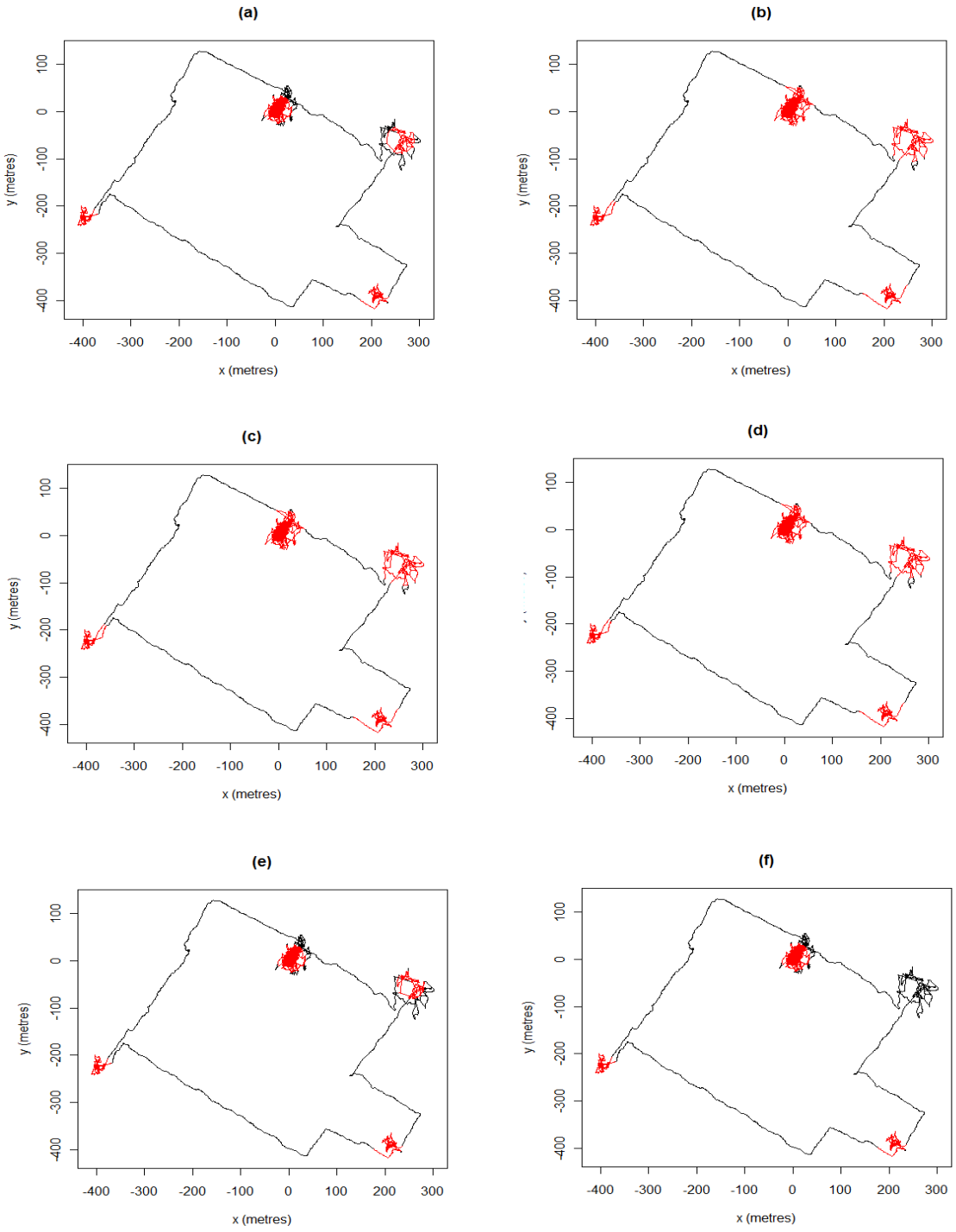

Figure 4.3: Plots of clusters for subject 1, time period 1, day 2, which are computed from various choices of parameter values:
(a) $\gamma = 0.2$, $s = 300$, $R = 30$, (b) $\gamma = 0.3$, $s = 300$, $R = 50$,
(c) $\gamma = 0.2$, $s = 300$, $R = 50$, (d) $\gamma = 0.2$, $s = 420$, $R = 50$,
(e) $\gamma = 0.1$, $s = 180$, $R = 30$, and (f) $\gamma = 0.1$, $s = 300$, $R = 30$.

Chapter 5

Results from the New Clustering Procedure

In this chapter, various results of the time-dependent clustering procedure are explored. Section 5.1 presents the results on the number of clusters/hotspots identified in each of the location time series. Section 5.2 presents the results on the proportion of time spent in the clusters, which gives an indication of how active and mobile the individual is. The area covered by the classical 95% ellipse, robust 95% ellipse and minimum spanning ellipse around the location points of each time series is discussed in Section 5.3 to give an idea of how far from the home the individual travelled and their lifespace. Examples are given in Section 5.4 to demonstrate the clusters identified for various data sets, the proportion of time spent in the clusters and the distance from these cluster centres to the home location.

5.1 Number of identified clusters

An aspect of an individual's mobility is the number of clusters/hotspots formed by their movements in their daily lives. If an individual has many clusters in their

daily recorded travel paths, this indicates that the individual is mobile enough to get around to those various locations. On the other hand, if an individual remains at their home location most days, this may imply they may have limited mobility which is restricting them to remain at their home location. This is a general guideline, as the recorded location points indicate where the individual actually went rather than what the individual is capable of doing.

The number of clusters/hotspots is considered to be the number of unique locations at which the individual spent an extended period of time. The amount of time required to be spent at each location is a quantity to be determined by the researchers, which should be a reasonable value such at five to thirty minutes. Therefore, if an individual's route for a particular day was home, school, work, home, grocery store, home, the number of clusters is considered to be 4 (home, school, work, grocery store) since the home location is only counted once even though it is visited multiple times in the day.

Figure 5.1 displays the number of unique locations of identified clusters for subject 1 for all available recorded days. The vertical lines go from zero up to the number of unique clusters identified by the new clustering procedure presented in Chapter 4. From Figure 5.1, one can see that 4 unique hotspots have been identified for day 1, whereas 6 hotspots have been identified for day 5 and 9 for day 9. In the algorithm, we set $t_d = 1$ second, $s = 5$ minutes, $r = 0.5$, $R = 30$, and $\gamma = 0.2$.

In Figure 5.2, the number of identified clusters for all recorded data is presented in a boxplot separated by the subjects. The two time periods have been combined for each subject and results are displayed for all participants. It can be seen that there is more variability in the number of clusters identified between the days for certain subjects in comparison to others. For instance, subject 1 has between 1 and 9 clusters identified on various days, whereas subject 13 has between 1 and 3 clusters identified

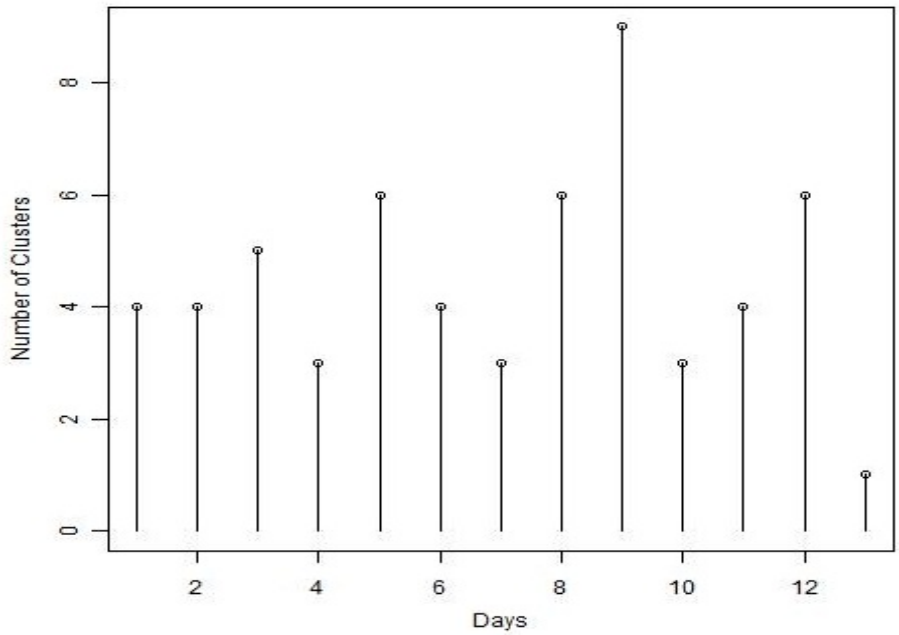

Figure 5.1: Number of clusters identified by procedure for subject 1.

on the various days of recorded data. Based on Figure 5.2, it looks as if the median number of clusters identified for each participants is not the same. However, more rigorous testing needs be done in order to determine whether the median number of identified clusters is the same between all the subjects.

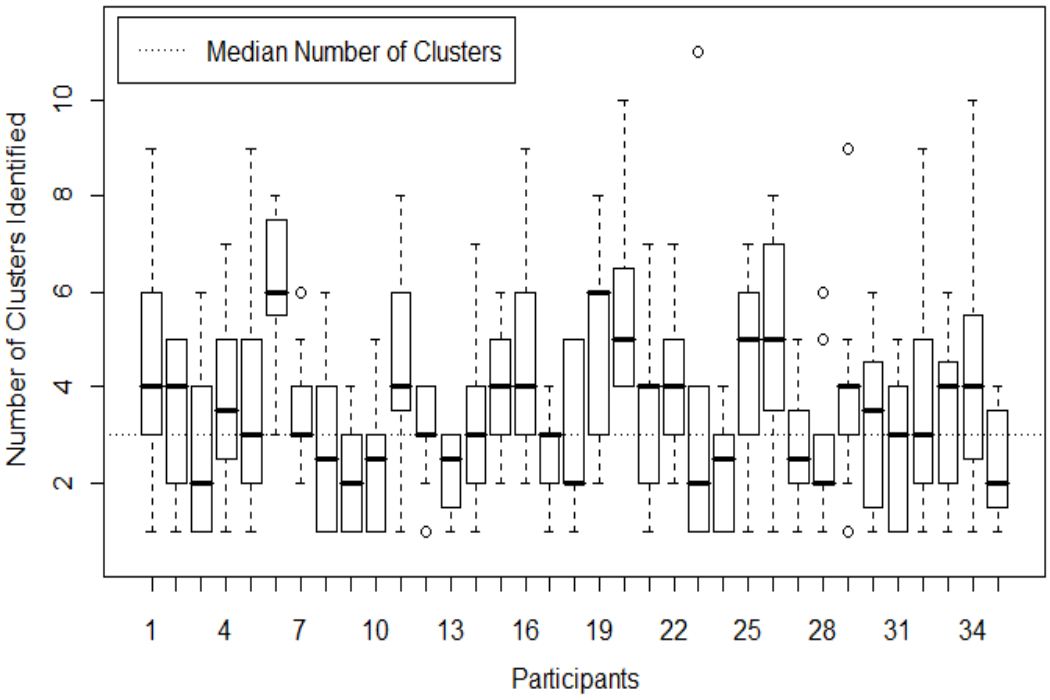

Figure 5.2: Boxplot of the number of clusters identified.

Table 5.1 displays the median number of identified unique clusters for each subject in the study for the two time periods. The median number of clusters ranged from 2 to 6 for these given subjects. The Kruskal-Wallis rank sum test was performed to test whether the median number of unique clusters was the same for each of the 35 subjects. The test resulted in a Kruskal-Wallis chi-squared test statistic of 95.9414 with 34 degrees of freedom and a p-value of 8.158×10^{-8}. Therefore, the median number of clusters is not the same for all of the subjects.

Table 5.1: Median number of clusters identified for each subject

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Median Number of Clusters	4	4	2	3.5	3	6	3	2.5	2	2.5	4	3
Subject	13	14	15	16	17	18	19	20	21	22	23	24
Median Number of Clusters	2.5	3	4	4	3	2	6	5	4	4	2	2.5
Subject	25	26	27	28	29	30	31	32	33	34	35	
Median Number of Clusters	5	5	2.5	2	4	3.5	3	3	4	4	2	

5.2 Proportion of time spent in clusters

Another important aspect of an individual's mobility is the amount of time spent inside or equivalently outside of the clusters. This is due to the fact that if the individual's location is not located in a cluster, he/she must be moving between the clusters demonstrating the ability to be mobile. If an individual spends a large proportion of the day not in any of the identified clusters, this individual would be considered to be very mobile in the sense that they were able to be on the move much of the day. There is no implication of their physical activity as the time spent outside of the clusters may have been walking, running, cycling, on a bus, in a car, etc. On the other hand, if an individual remains at home all day, they would not be considered mobile that day.

Table 5.2 displays the median proportion of time the individual spends in the identified clusters in comparison to the time the series spans. The two time periods have been combined in this table. It looks as if the median proportion of time spent in the hotspots are not the same for each of the subjects. This can be seen further from the boxplot presented in Figure 5.3. From Figure 5.3 it can be seen that some proportions go up to 100% and some go as low as approximately 10%. To test the

hypothesis that the proportions are the same for each subject, the Pearson's chi-squared test was performed. The test statistic was 3618.470 on 34 degrees of freedom and the resulting p-value is 2.2×10^{-16}, meaning that the proportion of time in the clusters is not the same for all subjects.

Table 5.2: Median proportion of time spent in the identified clusters for each subject

Subject	1	2	3	4	5	6	7	8	9
Proportion	0.628	0.820	0.864	0.799	0.841	0.779	0.544	0.866	0.881
Subject	10	11	12	13	14	15	16	17	18
Proportion	0.926	0.763	0.482	0.950	0.861	0.450	0.661	0.740	0.945
Subject	19	20	21	22	23	24	25	26	27
Proportion	0.822	0.678	0.725	0.841	0.858	0.441	0.691	0.651	0.577
Subject	28	29	30	31	32	33	34	35	
Proportion	0.464	0 .654	0.751	0.870	0.657	0.598	0.708	0.8210	

5.3 Ellipse construction and lifespace

Another aspect of an individual's movements that can give insight into his/her mobility is the area of a $100(1-\alpha)\%$ ellipse, which contains $100(1-\alpha)\%$ of the points. The area of the individual ellipses over each cluster can be computed to give an idea of the size of the clusters. The area of a $100(1-\alpha)\%$ ellipse for all of the location points in the time series can be computed to give insight into the overall space the individual occupied on a given day. If this area is small, it implies the individual did not venture far from the home location meaning the individual may not have the mobility capability to go far from the home. On the other hand, if this area is large, it implies the individual was able to go far from the home location implying they are mobile enough to move around quite a bit throughout the day.

The area of these ellipses may be calculated as $a \times b \times \pi$, where a and b are the

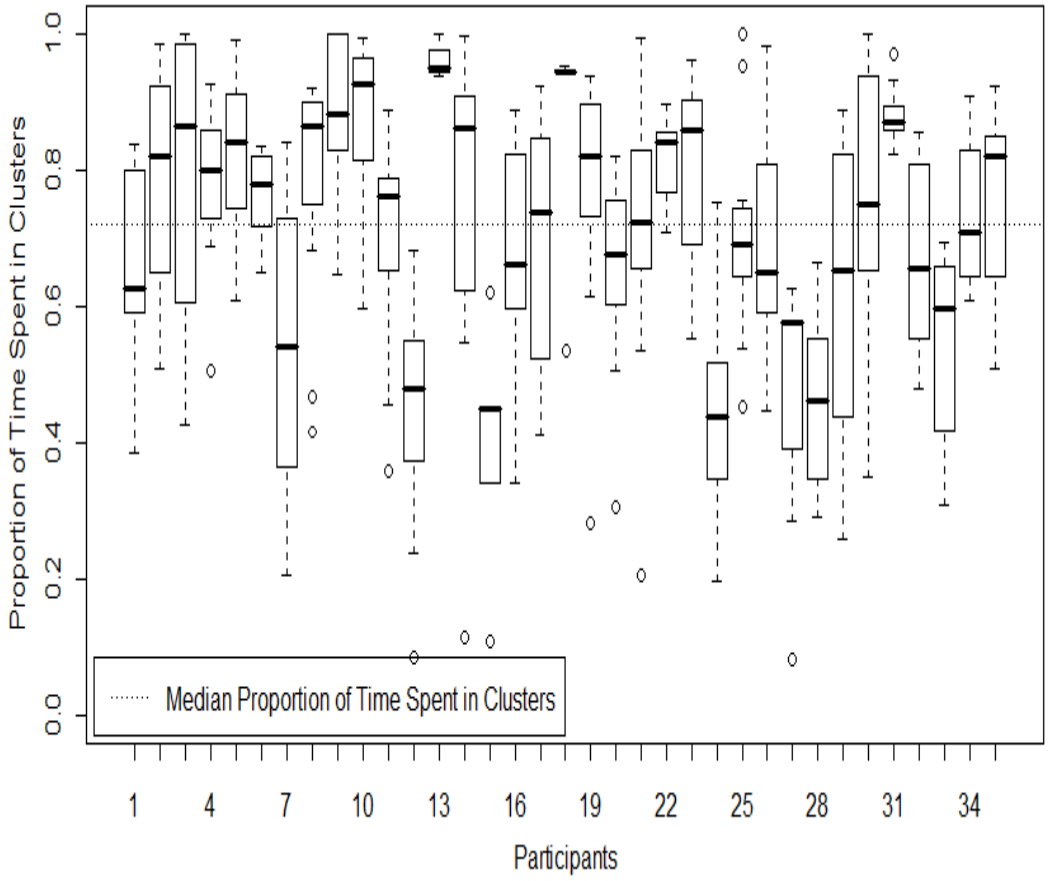

Figure 5.3: Proportion of time spent in the identified clusters for all subjects.

lengths of the major and minor axis of the ellipse. The lengths, a and b may be calculated as follows:

1. Calculate the centre point of the ellipse, (μ_x, μ_y).

2. Calculate the covariance structure of the data, $\mathbf{V}$.

3. Calculate the eigenvalues, λ_1 and λ_2, and the eigenvectors, e_1 and e_2, of the covariance matrix, $\mathbf{V}$.

4. Compute the values $z_i = \sqrt{c_q \times \lambda_i} \times e_i$ for $i = 1, 2$, where c_q is the $(1-\alpha)^{th}$ quantile of the computed distances, $(x_i - \mu_x, y_i - \mu_y)'\mathbf{V}^{-1}(x_i - \mu_x, y_i - \mu_y)$. Note: the axes of the ellipse may be plotted as two lines between the pairs of points $(\mu_x - z_{1_x}, \mu_y - z_{1_y})$, $(\mu_x + z_{1_x}, \mu_y + z_{1_y})$ and $(\mu_x - z_{2_x}, \mu_y - z_{2_y})$, $(\mu_x + z_{2_x}, \mu_y + z_{2_y})$, where z_{i_x} and z_{i_y} are the x and y coordinates of the point z_i for $i = 1, 2$.

5. Compute half of the length of the axes, $d_{z_i} = \sqrt{(z_{i_x})^2 + (z_{i_y})^2}$, for $i = 1, 2$. Note: $a = \min(d_{z_i})$ and $b = \max(d_{z_i})$ as in Figure 5.4.

Although the construction of the ellipse is fixed, there are various ellipses that can be constructed. This is due the fact that there is more than one way to obtain the centre point and covariance structure of the ellipse. For a classical ellipse the centre point would be (mean($\mathbf{x}$),mean($\mathbf{y}$)) and the covariance structure would be the classical covariance matrix using all the data points. However, there are robust methods that may be used to obtain the centre point and covariance structure for a robust ellipse. Rousseeuw and Van Driessen (1999) present a highly robust estimator of the multivariate location and scatter known as minimum covariance determinant (MCD). The objective of MCD is to find h observations out of n whose classical covariance matrix has the lowest determinant. The MCD estimate of location is the average of the h points, and the MCD estimate of scatter is their covariance matrix.

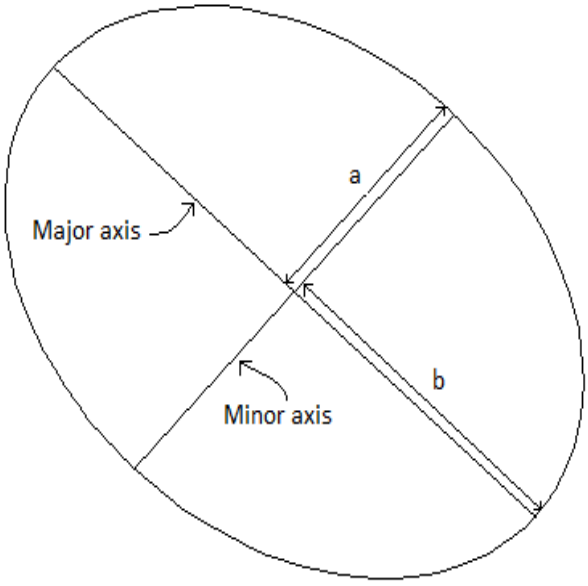

Figure 5.4: Ellipse with major and minor axes labelled.

The classical $100(1-\alpha)\%$ ellipse and the robust (MCD) $100(1-\alpha)\%$ ellipse work well for data sets that are "cloud" shaped. However, these methods do not fit the data well in other situations. This is due to the fact that both the classical and robust $100(1-\alpha)\%$ ellipse have an underlying assumption that the data is bivariate normal. Therefore, when the data have a shape that is similar to that of bivariate normal data, these two methods produce an ellipse that fits the data well. However, when the data deviates from a bivariate normal distribution, the classical $100(1-\alpha)\%$ ellipse and the robust $100(1-\alpha)\%$ ellipse do not fit the data well and therefore results in an undesirable lifespace ellipse and area. It should be noted that these constructions of the $100(1-\alpha)\%$ ellipse do not guarantee that all the points in the identified clusters are included. Furthermore, due to being based on the centre location and covariance structure, which is dependent on the shape of the data (cloud shaped, city block shape, etc.), the ellipse may cover a large area where there are no recorded data points.

Titterington (1978) presents another ellipse, the minimum spanning ellipse, also known as the ellipsoid hull or minimum area ellipse. This is the ellipse with minimum area such that all given points lie just inside or on the boundary of the ellipse. This may be more appropriate for data that do not follow a bivariate normal distribution, as it will preserve the shape of the data. It may also reduce the amount of area covered where no points are recorded due to being centred at a more appropriate value. One potential problem of this method is that it covers 100% of the points rather than $100(1-\alpha)\%$ as in the previously described ellipse constructions. The minimum spanning ellipse method does poorly if there are large outliers in the data set. For instance, if an individual turns off the GPS unit or loses signal, travels 10 kilometres away and then turns the unit back on or regains signal, we would want to include this true movement in the lifespace ellipse. However, if there is large noise

that causes the GPS to give readings 10 kilometres away for a short period of time, we would not want to include these location points in the ellipse as they are not representative of the individual's true location. In practice it is very difficult to tell the difference between these two situations. Therefore, if there are large outliers in the data, the minimum spanning ellipse will be greatly affected potentially resulting in a larger lifespace area than desired. To make the lifespace ellipse more robust in the sense of eliminating the wandering around inside a cluster, we reduce all points in each identified cluster to the centre point of the cluster, (median($\mathbf{x}$),median($\mathbf{y}$)), before computing the minimum spanning ellipse.

The area covered by a classical 95% ellipse for the entire time series for each subject is presented in Appendix D. The results for the classical 95% ellipse are presented in boxplots in Figure 5.5. In Figure 5.5 (a) it is difficult to see the individual boxes and the variation within the recorded area of the 95% ellipse for each subject. Figure 5.5 (b) is this same boxplot zoomed in individual boxes to gain better understanding of the recorded areas and the variation of the area for each subject. There appears to be large variations in the area of the 95% ellipses, which can be seen by the boxes for subject 9 and 16. Subject 9 has very little variation between the recorded areas, whereas the area of the 95% ellipses for subject 16 varies greatly between days as shown by the 25%-75% quantiles ranging from approximately 10 to 275 square kilometres.

5.4 Examples

In this section, a few examples are presented in detail. They show the results from the various measures of analysis done thus far in the thesis, such as total distance travelled, number of clusters identified, proportion of time in the cluster and distance

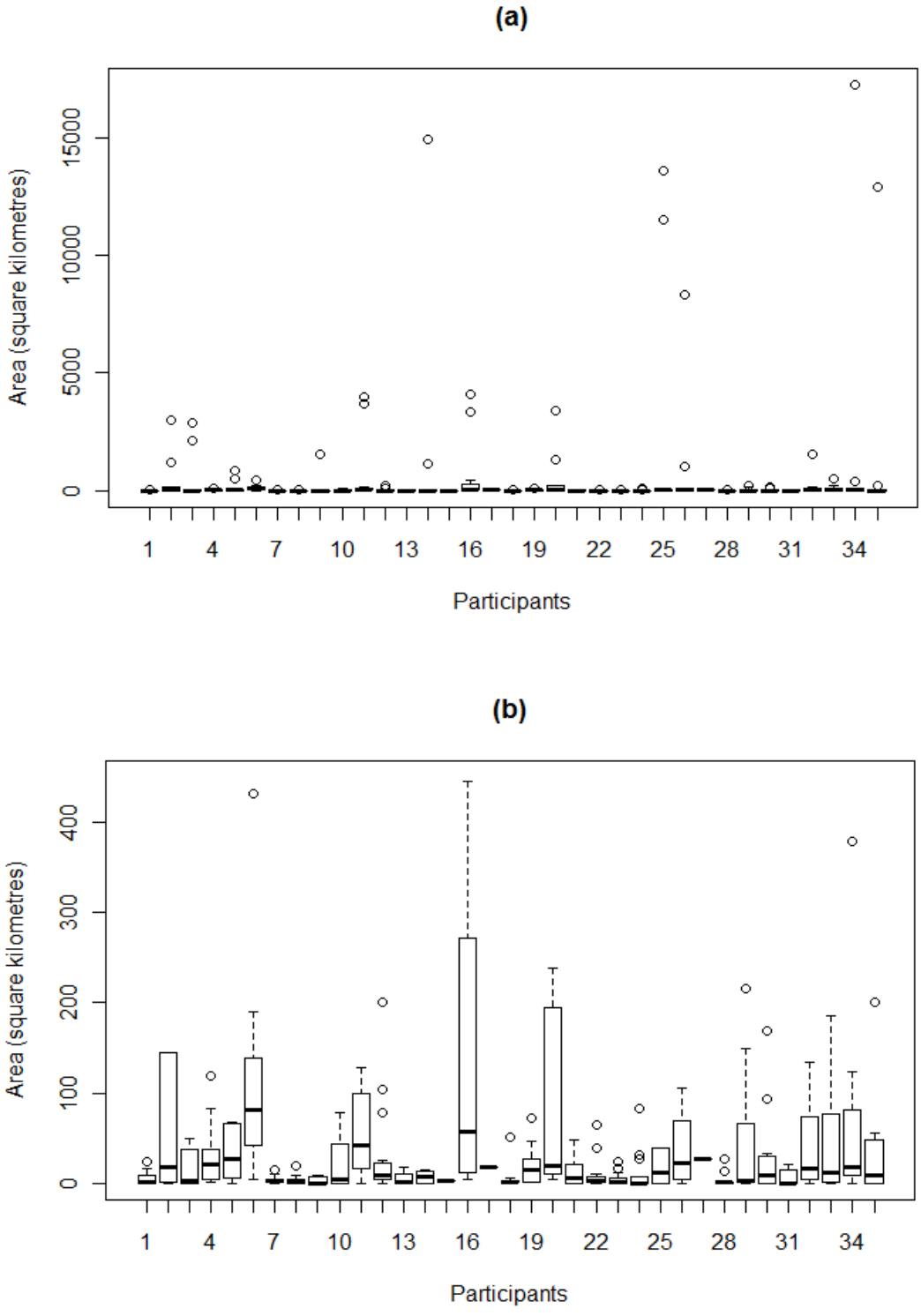

Figure 5.5: Boxplot of area covered by classical 95% ellipse:
(a) Total area covered, (b) Zoomed-in plot on boxes in boxplot.

the cluster centres are from home.

Example 5.1. *The data presented in this example are the recorded location points for subject 26, time period 1, day 4. The time series spans 10 hours, 38 minutes and 31 seconds. However, only 70.8% of these time points have recorded locations, making 29.2% of the series consist of interpolated data points. The total distance travelled for this individual on this given day is 21.1km. Figure 5.6 displays the location points for this individual's movements on this particular day. It can be seen that there are two areas in which the points are concentrated; one at the bottom left corner near* $(0, 0)$ *and a larger one near the upper right hand corner near* $(3200, 3000)$*. From the time-dependent clustering procedure, two clusters are in fact identified. The centres of these identified clusters are* $(6.2, -12.5m)$ *and* $(3211.0m, 3055.1m)$*, which are located 13.9m and 4421.2m from the home location of* $(0, 0)$*, respectively. It is clear that* $(0, 0)$ *is not the centre of an identified cluster, though it is fairly close to the centre of one identified cluster. It was found that 78.7% of the time series are part of one of the two clusters. Hence, this individual spent much of the day (approximately nine and a half hours) in one of these two places. Classical 95% ellipses have been placed around the two identified clusters in Figure 5.6 to indicate where one would expect the points to be located for each of the given clusters. Figure 5.7 displays the results from the k-means and trimmed k-means clustering algorithms. The parameters of* $k = 2$ *and* $\alpha = 0.1$ *were used. It is clear that the k-means algorithm does not perform well. On the other hand, the trimmed k-means algorithm appears to have identified appropriate locations for the clusters. The centres of the clusters are very similar for all three methods, but the number of points in the clusters vary quite a bit between the methods, as seen in Table 5.3. In the data set, there are* $n = 27142$ *observations for 38311 seconds. In the algorithm, we set* $t_d = 1$ *second,* $s = 300$ *seconds,* $r = 0.5$*,* $R = 30$*, and* $\gamma = 0.2$*.*

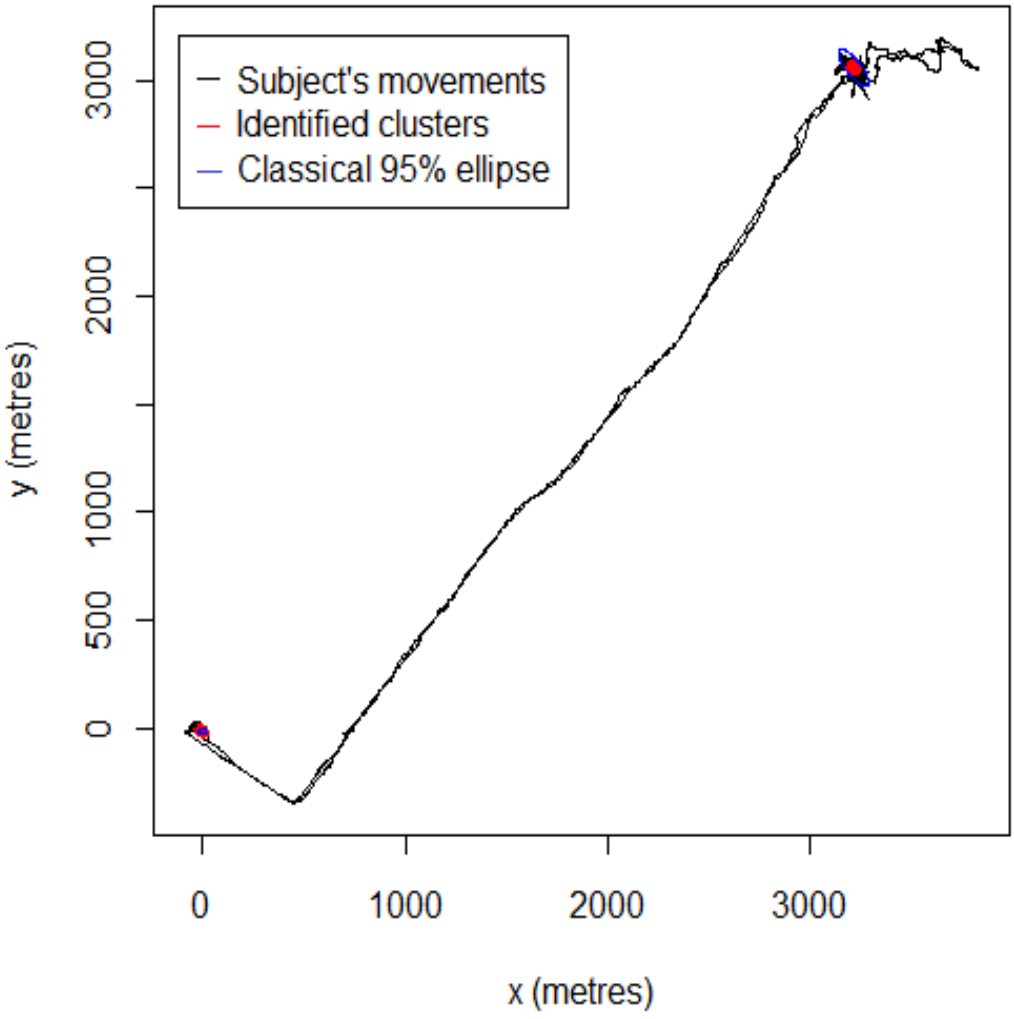

Figure 5.6: Plot of clustered time series for subject 26, time period 1, day 4.

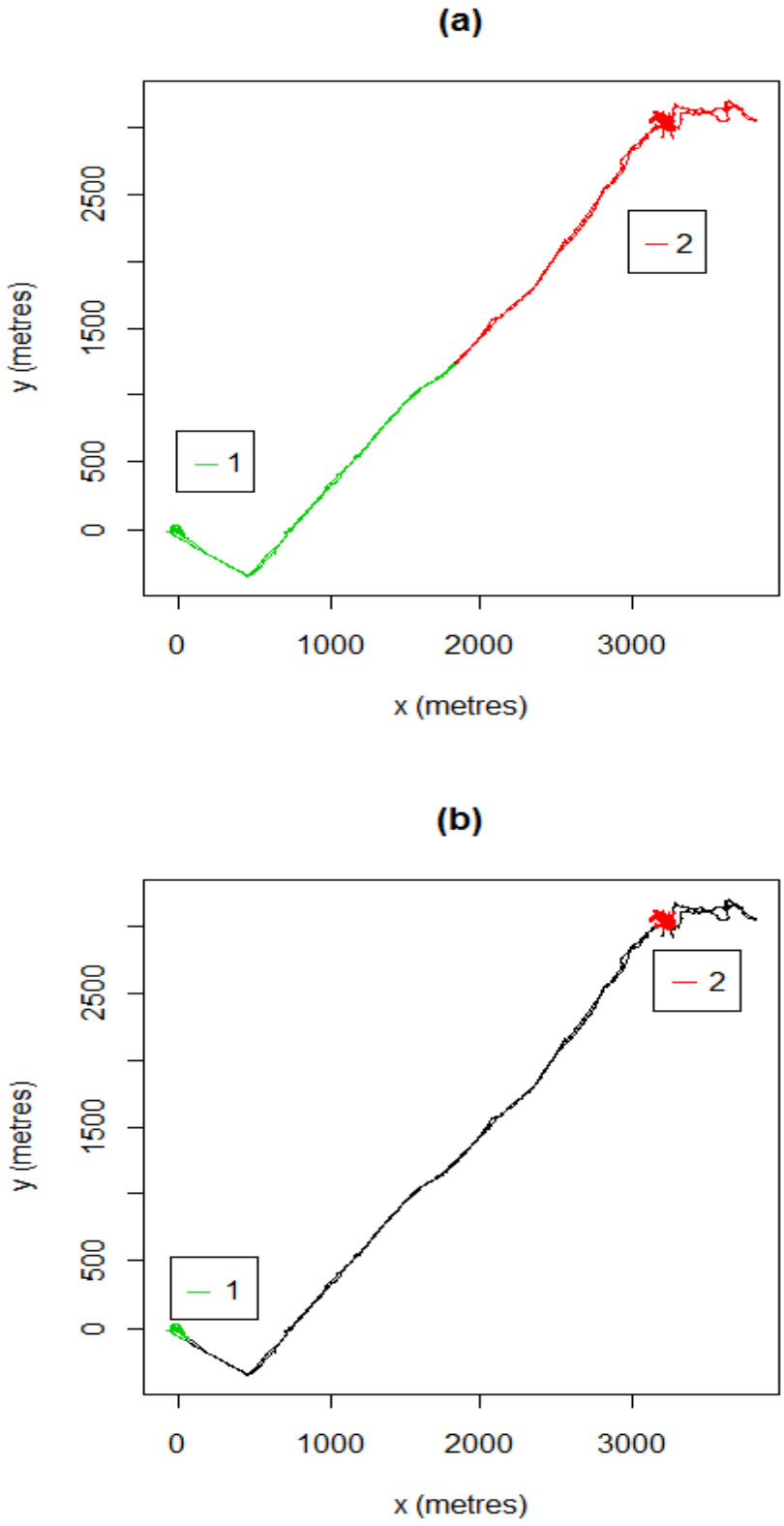

Figure 5.7: Plot of clustered time series for subject 26, time period 1, day 4: (a) k-means ($k = 2$) and (b) trimmed k-means ($k = 2, \alpha = 0.1$).

Table 5.3: Summary statistics for subject 26, time period 1, day 4

	Cluster	x_c	y_c	# Points in Cluster	Distance between cluster and (0,0)
New method	1	6.2	- 12.5	27112	13.9
$\gamma = 0.2$	2	3211.0	3055.1	2282	4421.2
Trimmed k-means	1	6.2	-12.5	27669	14.0
$\alpha = 0.1$ $k = 2$	2	3214.3	3049.3	6811	4430.6
k-means	1	6.5	-12.5	29086	14.0
$k = 2$	2	3215.0	3046.9	9225	4429.4
	Entire Series (4.4 km^2)	8.5	-10.6	47873	13.6

Example 5.1 gave results for one day of recorded data for subject 26. Since one cannot get an accurate idea of lifespace based on one example, let us examine another subject's daily location in detail.

Example 5.2. *The data in this example is from subject 12, time period 1, day 7. This individual's recordings spanned a total of 13 hours, 17 minutes and 53 seconds. Unfortunately, only 48.7% of the time series has recorded location points, meaning that over half of the time series has been filled in by the linear interpolation algorithm. We will continue with the analysis even though there is so much missing data. However, we should be aware of the low quality of data with so much of the series missing.*

Figure 5.8 displays the interpolated series for this particular individual. The total distance travelled by the individual on this day was 26km, which seems to be overly high considering it appears as if the subject does not travel far beyond their home.

This demonstrates the effect measurement errors have on the total distance travelled in a day. Although the majority of the points are quite close to one another, it is clear that there are some location points that appear to be sporadic and further out from the main group of points, which is contributing to a potentially inaccurately high total distance measurement. The time-dependent algorithm identifies one cluster with a center location of $(-3.0m, 205.0m)$*. Again, the identified cluster center is not at the assumed home location of* $(0, 0)$*, but rather 205.0m away. Although this is not extremely far away from the assumed home location, it does imply that not all individuals turn on their devices in the center location of their main activities within their home or that the GPS started with a poor fix. In this case it appears the first few data points in the series are distanced from the majority of the points which could be due to the individual turning on the device and then instantly moving toward the identified cluster or due to measurement error. This does identify a pitfall in using* $(0, 0)$ *as the home location in some data sets, but* $(0, 0)$ *is still a reasonable assumption for the home location.*

From the analysis, it is found that 81.5% of the data points are located in the identified cluster. This percentage is lower than one may expect and is most likely due to measurement errors that occur in this time series. The area covered by the ellipse is quite small implying the individual was not very mobile on this particular day. The classical and robust 95% ellipses, and the minimum spanning ellipse for the entire series has been placed over the time series in Figures 5.8 (a)-(c), respectively. From Figures 5.8 (a) and (b) it can be seen that both the classical and robust 95% ellipse, respectively, fit the data well. Figure 5.8 (c) displays the minimum spanning ellipse for the time series with points in the cluster reduced to the central point, which results in an area larger than one would desire considering the majority of the points are located in the cluster. In the data set, there are $n = 23303$ *observations for 47873*

seconds. In the algorithm, we set $t_d = 1$ *second,* $s = 300$ *seconds,* $r = 0.5$*,* $R = 30$*, and* $\gamma = 0.2$*.*

Table 5.4: Summary statistics for subject 12, time period 1, day 7

	x_c	y_c	# of Points	Distance between cluster and (0,0)
Cluster 1	-12.5	212.1	17786	212.5
Entire Series	-9.4	213.5	47873	213.7

Example 5.3. *This example displays the route travelled by subject 1 on day 2 of the first recorded time period. In Figures 5.9 (a)-(c) the classical 95% ellipse, robust 95% ellipse and minimum spanning ellipse for the entire time series have been placed over the recorded location points, respectively. These ellipses may be considered as the lifespace for this individual on this particular recorded day, since this is the area in which the individual spent the majority of their time on this particular day. From Figure 5.9 it can be seen that the classical and robust 95% ellipses are very similar. Both appear to give undesirable ellipses, as they cover a lot of space that has no recorded location points. This is due to the centre point of these ellipses being pulled further right due to many points being located there. The minimum spanning ellipse fits the data better in this situation, as it fits the data tightly. Note that all points located in the identified clusters have been set to each cluster's central value before fitting the minimum spanning ellipse. From Figure 5.9 (c), it can be seen that some points located near (-400,-220) have been cut out. This is due to those points being located in that cluster and therefore having been set to that cluster's central value. In this situation, the minimum spanning ellipse with all points in each identified cluster reduced to the central value of each cluster is considered to be the most desirable lifespace ellipse.*

Example 5.4. *This example displays the travelled path of subject 2, time period 2, day 1. The classical and robust 95% ellipses, as well as the minimum spanning ellipse*

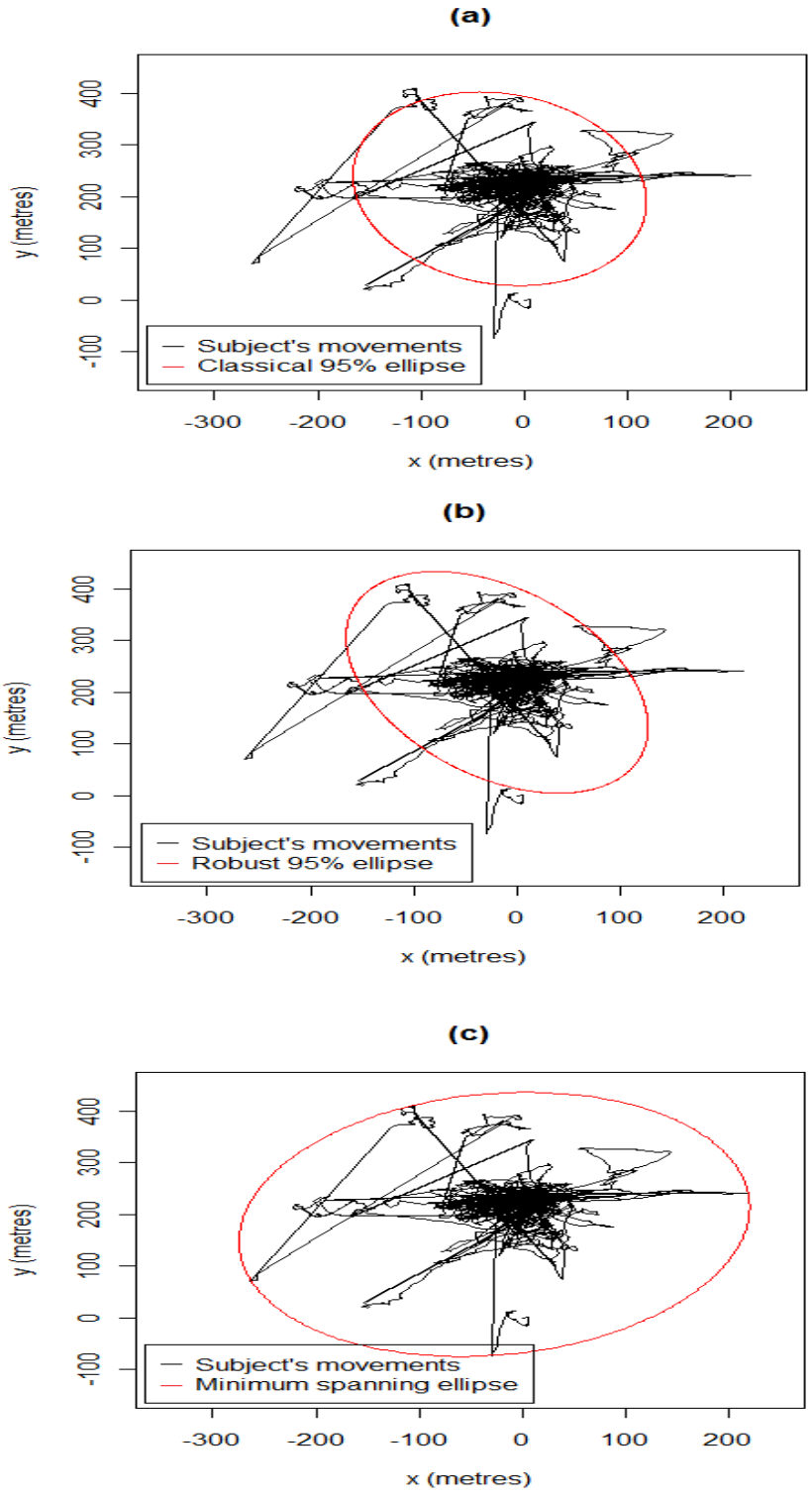

Figure 5.8: Plot of time series for subject 12, time period 1, day 7 in black with:
(a) classical 95% ellipse in red,
(b) robust 95% ellipse in red, and
(c) minimum spanning ellipse in red.

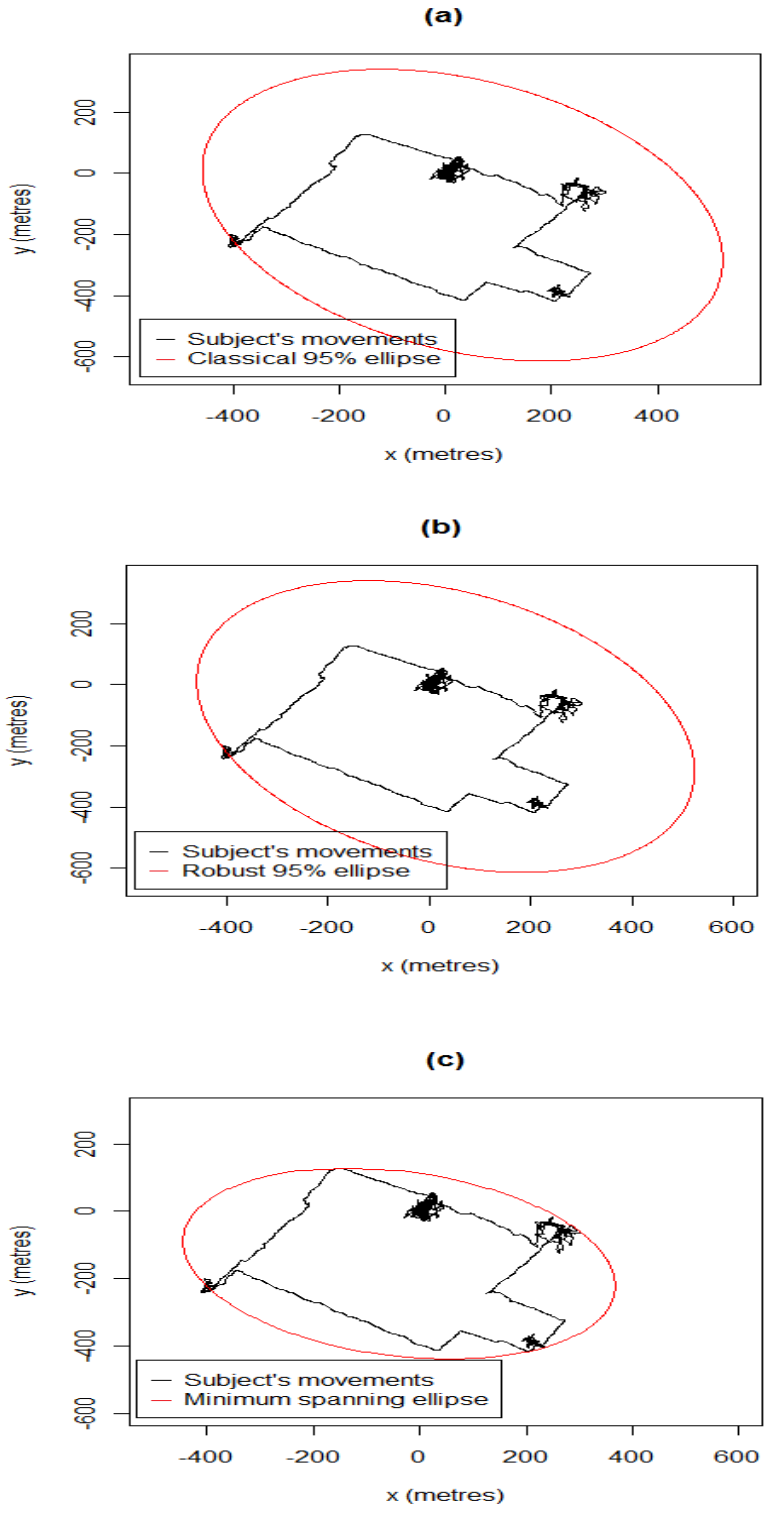

Figure 5.9: Plot of time series for subject 1, time period 1, day 2 in black with:
(a) classical 95% ellipse in red,
(b) robust 95% ellipse in red, and
(c) minimum spanning ellipse in red.

with points in each identified cluster reduced to each cluster's central value for the entire time series has been placed over the recorded location points in Figures 5.10 (a)-(c), respectively. As in Example 5.3, the ellipse may be considered as the lifespace of subject 2 for day 1 of the second recorded time period. It is clear that the classical and robust 95% ellipses do not fit the data overly well. They both extend quite far beyond any recorded location point. This is due to the fact that the individual travelled quite far on this particular day with most points located in the ends of the trajectory causing the variance in that direction to be large, which in turn causes the ellipse to be stretched out with a large major axis. However, the minimum spanning ellipse appears to fit the data fairly well in this situation. It encompasses the data tightly without covering a lot of area with no recorded data points.

As lifespace is a measure of the area an individual covers in their daily lives, it may be more appropriate to look at data for an entire week rather than a single day. This is due to the fact that an individual may stay at home one day and then travel quite a bit another day.

Example 5.5. *This example displays the travelled paths of subject 2 in time period 1. The data points from all days available for subject 2 in time period 1 are put together to gain a better idea of the lifespace for this individual. Figure 5.11 (a) displays the time series in black and the classical 95% ellipse in red, Figure 5.11 (b) displays the time series in black with the robust 95% ellipse in red, and Figure 5.11 (c) displays the time series in black with the minimum spanning ellipse in red. Note that when computing the minimum spanning ellipse, the points within each identified cluster have been set to the clusters' central values,* $(median(\mathbf{x}), median(\mathbf{y}))$*. The function cov.rob() was used in R to compute the robust estimates from the MCD algorithm with* $\lfloor 0.95 * n \rfloor$ *good points, and the function ellipsoidhull() was used in R to compute the minimum spanning ellipse. It is clear that the three methods result in very different*

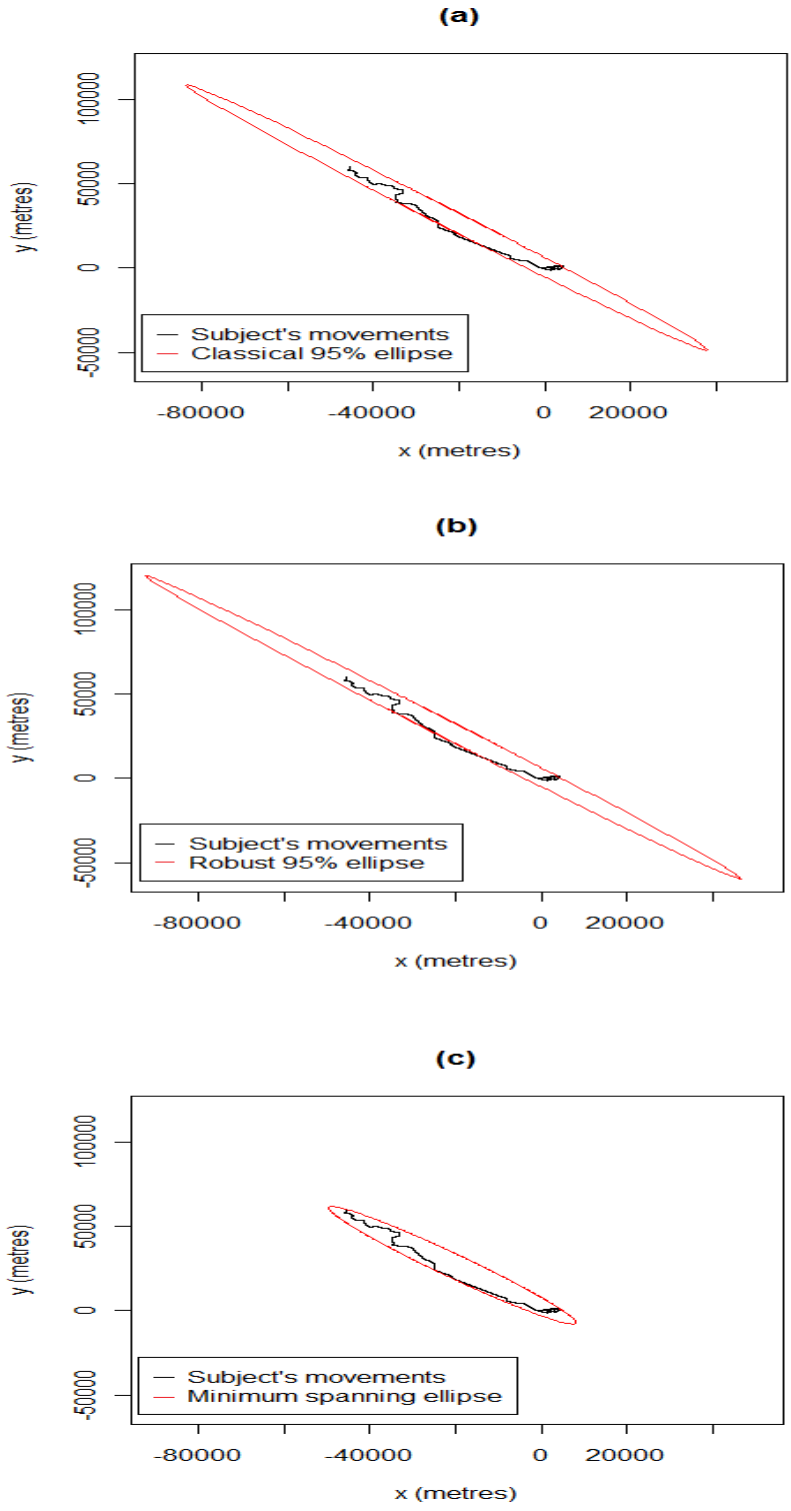

Figure 5.10: Plot of time series for subject 2, time period 2, day 1 in black with:
(a) classical 95% ellipse in red,
(b) robust 95% ellipse in red, and
(c) minimum spanning ellipse in red.

ellipses. It is important to note that the shape and size of the robust ellipse greatly depends on the number of "good points" used in determining the center location and covariance structure in the MCD algorithm. Although the total area of the classical and robust 95% ellipses are far smaller than that of the minimum spanning ellipse, they do not capture the shape of the data well.

In this chapter, several results were presented based on the time-dependent clustering procedure presented in Chapter 4. Although the algorithm identifies the logical number of clusters for each time series, a serious issue of measurement error has been identified. Classical and robust 95% ellipses, as well as minimum spanning ellipses were presented. The classical and robust 95% ellipses work well in cases where the data roughly follows a bivariate normal distribution. When the data deviates from a bivariate normal distribution, the minimum spanning ellipse works best, as it captures the shape of the data unlike the classical and robust ellipses presented. In the following chapter, we will continue the discussion on the measurement error problem that arises with this type of data and give some insight on how to deal with such errors.

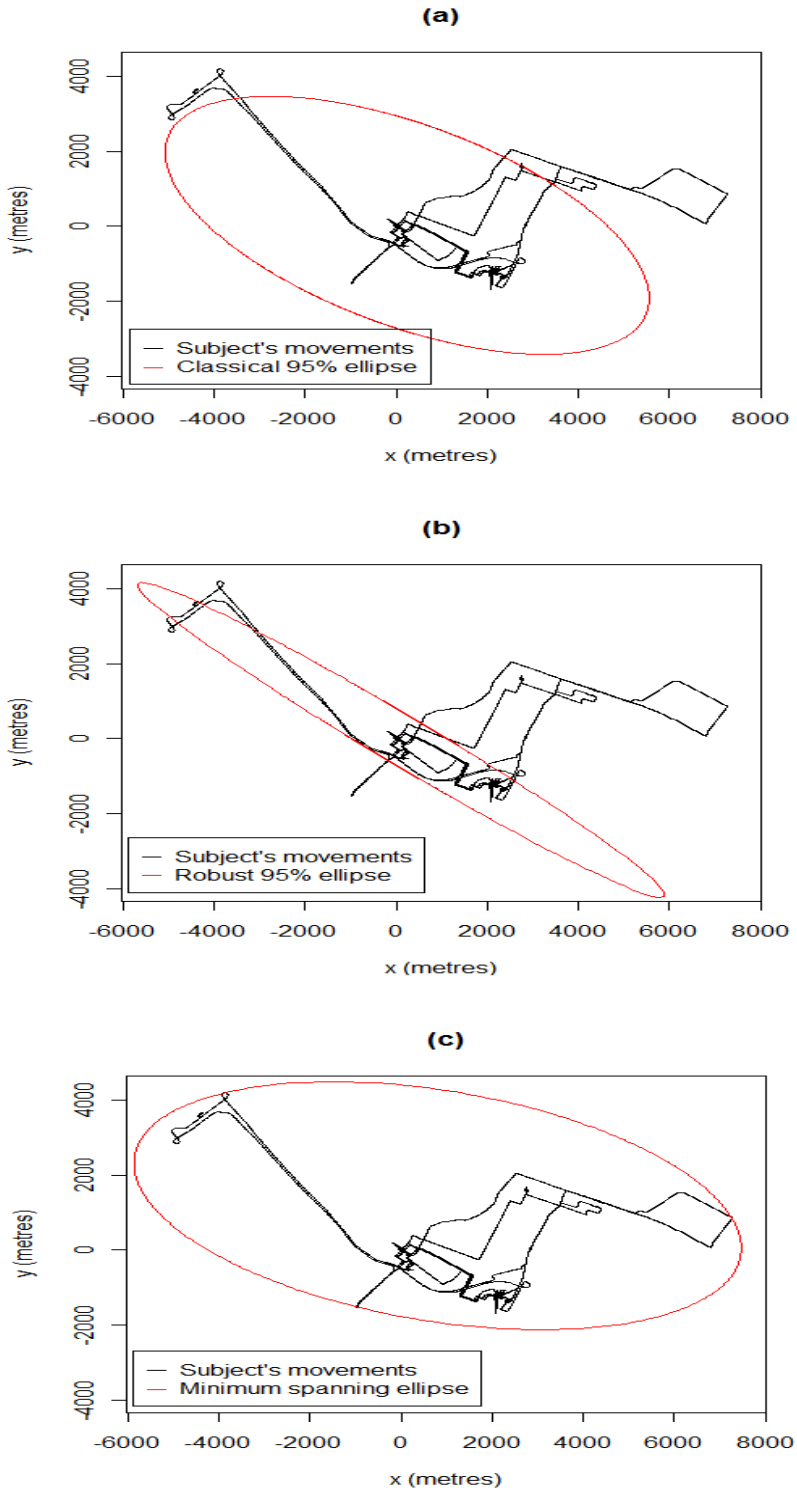

Figure 5.11: Plot of time series for subject 2, time period 1 time black with:
(a) classical 95% ellipse in red,
(b) robust 95% ellipse in red, and
(c) minimum spanning ellipse in red .

Chapter 6

Large Noise Detection and Smoothing Techniques

This chapter focuses on the noise within the time series of location points. All recorded location points have measurement error to varying degrees. We are interested in determining which points have a large amount of error causing inaccurate results. Section 6.1 gives insight into why identifying how much noise and where it is located within the series is important. Section 6.2 presents techniques to identify where the large noise is located within the series and Section 6.3 presents various methods to filter the time series to get more accurate results. Examples are given in Section 6.4 to demonstrate where the large noise is occurring within the chosen location time series, as well as how the series appear after filtering techniques have been applied. In Section 6.5, the distance travelled is recalculated using the filtered time series and considered as an upper bound on the true total distance travelled. A lower bound on the total distance travelled will also be presented.

6.1 Noise and its implications

As has been mentioned, there is an element of noise amongst the location time series data presented and analyzed throughout this thesis. As discussed by Kaplan and Hegarty (2006), Masumoto (1992), Parkinson and Spilker (1996), Radi (2012), Tsui (2005), and Wells et. al (1978), noise within the series can occur for many reasons including but not limited to the following:

- There can be timing errors due to ionosphere and troposphere delays resulting from the signal slowing down as it passes through the atmosphere, as well as receiver clock errors due to the GPS receiving device having a built-in clock that is less accurate than the atomic clocks on the satellites. Furthermore, signal multipath errors occur when the signal is reflected off objects prior to reaching the GPS receiving device causing errors due to the signal's travel time being increased.

- There can be errors in the reported location due to orbital errors resulting from inaccuracies of the satellite's reported location. Also, buildings and terrain can block the signal causing errors in the reported location or potentially no reported location. Furthermore, location errors can arise from poor satellite geometry, which occurs when the satellites are located in a line or close together. Ideally, there are many satellites visible to the GPS receiving device that are located at wide angles relative to one another with no objects blocking the signal's path.

Masumoto (1992) describes that it is possible to have very noisy data when the individual is inside a building or underground, or when the GPS has a fix with only 2 satellites since a minimum of three are needed for accurate recordings. Furthermore, it was observed that the data is far noisier when the individual is in the same location for an extended period of time.

The large noise may significantly alter a few of the measures presented in previous chapters, such as the total distance travelled. If there is large measurement error present in the time series, the calculated total distance is significantly higher than it should be.

If we are able to identify which intervals have large noise, we may be able to clean the data set and achieve more accurate results. With a "clean" series, the distance measure will tend towards the true distance the individual travelled on a particular day rather than an overestimate of this distance. Furthermore, we will not have to rely so heavily on robust measurements when identifying the clusters since the location points would all be fairly close to the accurate location points.

6.2 Large noise detection

The aim is to identify where the large amounts of noise within the time series are located in order for the series to be filtered or cleaned in some manner to obtain more accurate results. Many methods for identifying noise were explored and a few will be discussed in this thesis.

Through experimentation, various procedures were investigated to identify windows of time where high levels of noise are located. In all the procedures, a scrolling time window of length $l(w)$ is used and each window is determined as very noisy or not. Windows of time were used when finding large noise, as we are not overly concerned about one or two bad recorded points, but rather where large amounts of noise are located. The scrolling windows of time are non-overlapping and consecutive as shown in Figure 6.1.

The algorithms investigated are as follows:

1. Scroll through the time series with a moving window of length $l(w)$, where the

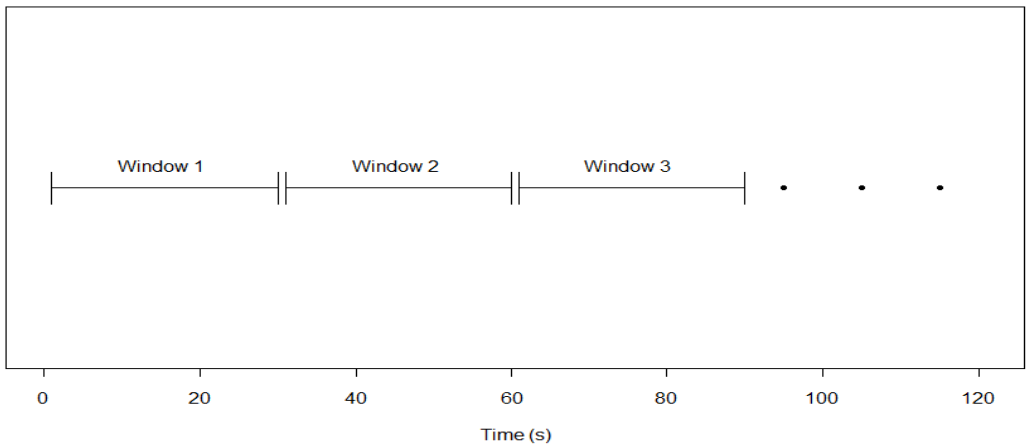

Figure 6.1: Non-overlapping windows of length $l(w) = 30s$ used to find large noise.

first window starts at the first data point in the series and the windows do not overlap. Therefore, the first window contains the location points for time points 1 through $l(w)$, the second window contains time points $l(w)+1$ through $2l(w)$ and so on.

2. (a) Compute the acceleration (or second difference) in each dimension of the time series at each time point. Denote the acceleration in the x direction by $a_{x_t} = (x_{t+2} - x_{t+1}) - (x_{t+1} - x_t)$ and the acceleration in the y direction by $a_{y_t} = (y_{t+2} - y_{t+1}) - (y_{t+1} - y_t)$ for $t = 1, 2, \ldots, n-2$, where n is the length of the interpolated series. Compute the amplitude of acceleration, a_t, at each time point within the time series, where $a_t = \sqrt{a_{x_t}^2 + a_{y_t}^2}$ for $t = 1, 2, \ldots, n-2$. For each window w_i, compute the average amplitude of acceleration and standard deviation of the amplitude of acceleration, denoted by $\bar{a}_{w_i}$ and $s_{a_{w_i}}$, respectively. For instance, $\bar{a}_{w_1} = \frac{1}{l(w)} \sum_{j=1}^{l(w)} a_j$, $\bar{a}_{w_2} = \frac{1}{l(w)} \sum_{j=l(w)+1}^{2l(w)} a_j$, $s_{a_{w_1}} = \sqrt{Var(a_1, \ldots, a_{l(w)})}$, $s_{a_{w_2}} = \sqrt{Var(a_{l(w)+1}, \ldots, a_{2l(w)})}$, etc.

   (b) Compute the distance between each location point in consecutive time. The Euclidean distance will be used and is calculated as $d_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}$ for $t = 2, \ldots, n$. For each window w_i, compute the mean of the distances and the standard deviation of the distances for the time points in the window, denoted by $\bar{d}_{w_i}$ and $s_{d_{w_i}}$, respectively. For example, window 1 would have $\bar{d}_{w_1} = \frac{1}{l(w)} \sum_{j=1}^{l(w)} d_j$ and $s_{d_{w_1}} = \sqrt{Var(d_1, \ldots, d_{l(w)})}$, window 2 would have $\bar{d}_{w_2} = \frac{1}{l(w)} \sum_{j=l(w)+1}^{2l(w)} d_j$ and $s_{d_{w_2}} = \sqrt{Var(d_{l(w)+1}, \ldots, d_{2l(w)})}$, etc.

3. If the computed measure is greater than a given cut-off value, κ, consider window w_i to contain large amounts of noise. Otherwise, consider window w_i to have

small noise. The measures for the various procedures are:

(i) Average amplitude of acceleration, $\bar{a}_{w_i}$.

(ii) Standard deviation of the amplitude of acceleration, $s_{a_{w_i}}$.

(iii) Ratio of standard deviation to mean amplitude of acceleration, $s_{a_{w_i}}/\bar{a}_{w_i}$.

(iv) Standard deviation of distance, $s_{d_{w_i}}$.

(v) Ratio of standard deviation to mean distance, $s_{d_{w_i}}/\bar{d}_{w_i}$.

The average amplitude of acceleration was investigated since high accelerations may imply large noise as a stationary person or someone walking, driving, etc. will not have a high average acceleration. Even if the individual is in a car starting to move after a red light, the acceleration will be large when the individual begins to move but the acceleration will tend to zero fairly quickly. Therefore, it is unlikely that natural movements in location result in high average acceleration over a window of time. This method works fairly well at identifying where large jumps in the location points occur and which windows in time are quite noisy. This method is determined to be effective purely by looking at which windows were identified and observing that large jumps in location were identified. However, this method does not address a few issues. It will not identify windows that have mostly/all interpolated points, as they are equally spaced along a straight line. This means that if the location points jump from one spot to another with a time lag and no recorded data in between, the interpolated data will not be identified as noise and therefore one cannot distinguish whether that jump in location is real or an error.

The standard deviation of distance was investigated due to the fact that it is expected that over small windows of time, in general, an individual will have roughly the same distance travelled each second. The goal was to identify windows with large variations of distance as this would imply possible unrealistic jumps in location.

Similar to the standard deviation of distance method, the method using the standard deviation of the amplitude of acceleration was looked into as it is assumed that over a short window of time the amplitude of acceleration of the individual's movements would be approximately zero. Thus, a large standard deviation of amplitude of acceleration would imply time points with large accelerations are likely caused by large unrealistic jumps in location.

The methods using the ratio of the standard deviation to mean of the amplitude of acceleration and distance are based on statistical quality control theory. Burr (1979), Jamieson (1982), Montgomery (1985) and Pyzdek (1989) all discuss how statistical quality control theory aims to identify large variations in the data.

For all updated results and filtering methods, the average amplitude of acceleration, $\bar{a}_{w_i}$, will be used as it is a simple measure to calculate and heuristically appears to identify appropriate windows of time as very noisy.

6.3 Smoothing techniques

Several methods have been proposed in Section 6.2 to identify where large noise in the time series noise is located. In this section, various filtering techniques are investigated and their advantages and disadvantages are discussed.

6.3.1 Moving average

In time series, a moving average is commonly used to smooth out short-term fluctuations and small amounts of noise in addition to preserving long-term trends. A moving average may be viewed as a low-pass filter in signal processing. The mean is taken from an equal number of data points on either side of the central value. The number of points on either side of the central value may be adjusted to fit the need of the situ-

ation. As shown by Shumway and Stoffer (2006), to calculate a moving average of the time series $x_1, x_2, \ldots, x_n$ with k points on either side of the central value, the following formula is used: $\hat{x}_i = (x_{i-k}+x_{i-k+1}+\ldots+x_{i-1}+x_i+x_{i+1}+\ldots+x_{i+k-1}+x_{i+k})/(2k+1)$.

A few advantages of this method include the fact that the procedure is easily applied to large time series, it is a well-known and easily understood filtering technique in time series analysis and the smoothing parameter, $2k+1$, may be adjusted. However, there are some large drawbacks to this method in regards to location time series, including the fact that if the moving average is applied to the entire time series, the distinctive shape of the individual's movements are lost as the corners are smoothed out. For instance, if an individual walked around city blocks for a portion of the day, their movements would include some sharp corners where they changed direction. The moving average would smooth these corners and round out the movements and therefore change the shape of the location points. If the moving average is applied to only the windows indicated as being very noisy, there are still some issues that may arise. Since the moving average takes an average of neighbouring points, even a single large measurement error would skew all the neighbouring points in the moving average.

6.3.2 Elimination of high accelerations

Due to the fact that unrealistically high amplitudes of acceleration represent large noise within the location time series, one filtering method that was considered involved eliminating the high accelerations. This method is implemented as follows:

1. Compute the second derivative of the data in both the x and y directions. Hence, compute the second differences $a_{x_t} = (x_{t+2} - x_{t+1}) - (x_{t+1} - x_t)$ and $a_{y_t} = (y_{t+2} - y_{t+1}) - (y_{t+1} - y_t)$ for $t = 1, 2, \ldots, n-2$ where n is the length of the interpolated series of location points.

2. Absolute value of accelerations in either the x or y direction that are above a predetermined cut-off value, η_a, are set to 0.

3. The new/filtered location points are determined by integrating the acceleration time series twice, using the first point in the velocity (first difference) and location series as the two constants of integration.

This method of eliminating points with high absolute value acceleration was investigated as one would expect the majority of the acceleration series to be around zero. However, this method does not work well on location time series. When this method is applied to the entire series, the shape of the individual's movements may be entirely lost. This is due to the fact that the measurement errors do not occur symmetrically. After integrating the new acceleration time series in both the x and y directions twice, the errors accumulate throughout the entire series and the shape of the location series may be lost. This can be seen in Figure 6.3 (c) in Example 6.1.

6.3.3 Elimination of high velocities

This method is very similar to the previously described method. Instead of changing high accelerations, this method sets high velocities to zero. Again, this method is proposed as one expects the velocity to remain fairly constant and around zero for the majority of the series. Velocities that are too large to be realistic for an individual to be traveling at will be set to zero and the location points will be found as follows:

1. Compute the first derivative of the data in both the x and y directions. Hence, compute the first differences $v_{x_t} = (x_{t+1} - x_t)$ and $v_{y_t} = (y_{t+1} - y_t)$ for $t = 1, 2, \ldots, n-1$.

2. Absolute value of velocities that are above a predetermined cut-off value, η_v, are set to zero.

3. The new/filtered location points are determined by integrating the velocity time series, using the first point in the location time series as the constant of integration.

When this method is applied to the entire series, the individual's movements may be shifted, but the shape is preserved unlike in the previous method when the high accelerations were set to zero. Due to the fact that the aim of filtering the time series is to smooth the very noisy parts of the data only and leave the general shape and true location points of the individual's daily path intact, this method does not work well for our time series of location points.

6.3.4 Trimmed means

It has been shown in the moving average method that taking an average of points in a given time span does not smooth the series in an appropriate manner. We will now attempt to filter the series using trimmed means.

This trimmed means method is performed using the following steps:

1. Consider the points $(x_{i-j}, y_{i-j}), \ldots, (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}), \ldots, (x_{i+j}, y_{i+j})$. When $i-j$ is outside of the time points of the window of consideration, use the points from the previous window; similarly, if $i+j$ is outside of the window, use the points from the next window to calculate the trimmed mean.

2. Calculate the $(1-2\beta)\times 100\%$ trimmed mean of the points $(x_{i-j}, y_{i-j}), \ldots, (x_{i+j}, y_{i+j})$ in each the x and y directions using a predetermined trimming parameter β, $0 < \beta < 0.5$. The trimmed means in the x and y directions are denoted by $\hat{x}_i$ and $\hat{y}_i$, respectively.

This method is able to handle large jumps in data by bringing those data points closer to the true location due to the trimmed mean not being affected by large jumps

in data. On the contrary, some of the disadvantages to this method of filtering include the fact that if the trimmed mean is applied to the entire time series, the distinctive shape of the individual's movements are lost as the corners are smoothed out. Note that filtering is done only on points in windows identified as very noisy. Therefore, if time point t_j is not in a very noisy window, $\hat{x}_{t_j} = x_{t_j}$.

Due to the fact that we know which windows in time are very noisy, only these identified windows need to be filtered. There is no need to filter the other windows in time, as they have small noise. The trimmed means method appears to work the best, in terms of large jumps and drifts in data location being moved to the main areas of activity or paths at that time, by looking at the resulting filtered series. This method was modified slightly to allow more filtering on noisier windows of time. This is therefore an adaptive method. The adjusted method is as follows:

1. Determine how noisy the window is. For a predetermined number of cut-off values, $0 < \kappa_1 < \ldots < \kappa_p$, if window w_i has the value $\kappa_l < \bar{a}_{w_i} \leq \kappa_{l+1}$, then mark window w_i as having noise level l.

2. Consider the points $(x_{i-j}, y_{i-j}), \ldots, (x_{i-1}, y_{i-1}), (x_i, y_i), (x_{i+1}, y_{i+1}), \ldots, (x_{i+j}, y_{i+j})$. When $i - j$ is outside of the time points of the window of consideration, use the points from the previous window; similarly, if $i + j$ is outside of the window, use the points from the next window to calculate the trimmed mean.

3. Calculate the $(1 - 2\beta_l) \times 100\%$ trimmed mean in each the x and y directions, where the trimming proportion, β_l, depends on the noise level l associated with window w_i. The larger the value of l is, the larger the trimming proportion β_l is.

6.4 Examples

We will now look at a few examples displaying which windows of time the various noise identification methods identify as being very noisy, as well as the filtered time series resulting from the various filtering procedures discussed.

Example 6.1. *Consider the time series of location points from subject 1, time period 1, day 2. Figure 6.2 (a) displays the unfiltered time series as a thin black line with the location points in the identified clusters in red. We can see that the individual travelled in what appears to be a city block pattern with 4 identified clusters. By glancing at the location points plotted, there appear to be no serious problems with noise in this time series as there are no places where there are points are visibly quite far away from the neighbouring points of the time series.*

Figures 6.2 (b)-(f) present the points located in the windows identified as being very noisy by the five identification methods: (b) average amplitude of acceleration method with $\kappa = 1.25$*, (c) standard deviation of distance method with* $\kappa = 2$*, (d) standard deviation of amplitude of acceleration method with* $\kappa = 2$*, (e) ratio of standard deviation to mean distance method with* $\kappa = 1.25$ *and (f) the ratio of standard deviation to mean amplitude of acceleration method with* $\kappa = 2$ *as a blue line. It can be seen that the methods with these particular* κ *values identify different windows as being very noisy. Furthermore, each method identifies a different number of windows. It may be noted that all the flagged windows occur at identified clusters. This is consistent with the fact that we believe much of the large noise in the series arises near the identified clusters.*

Figure 6.3 (a) presents the unfiltered time series as a thin black line with the location points identified as part of a cluster in red. Figures 6.3 (b)-(f) present the filtered series resulting from the five discussed smoothing techniques: (b) moving average with 21 points , (c) eliminating high accelerations with $\eta = 1.25$*, (d) eliminating high ve-*

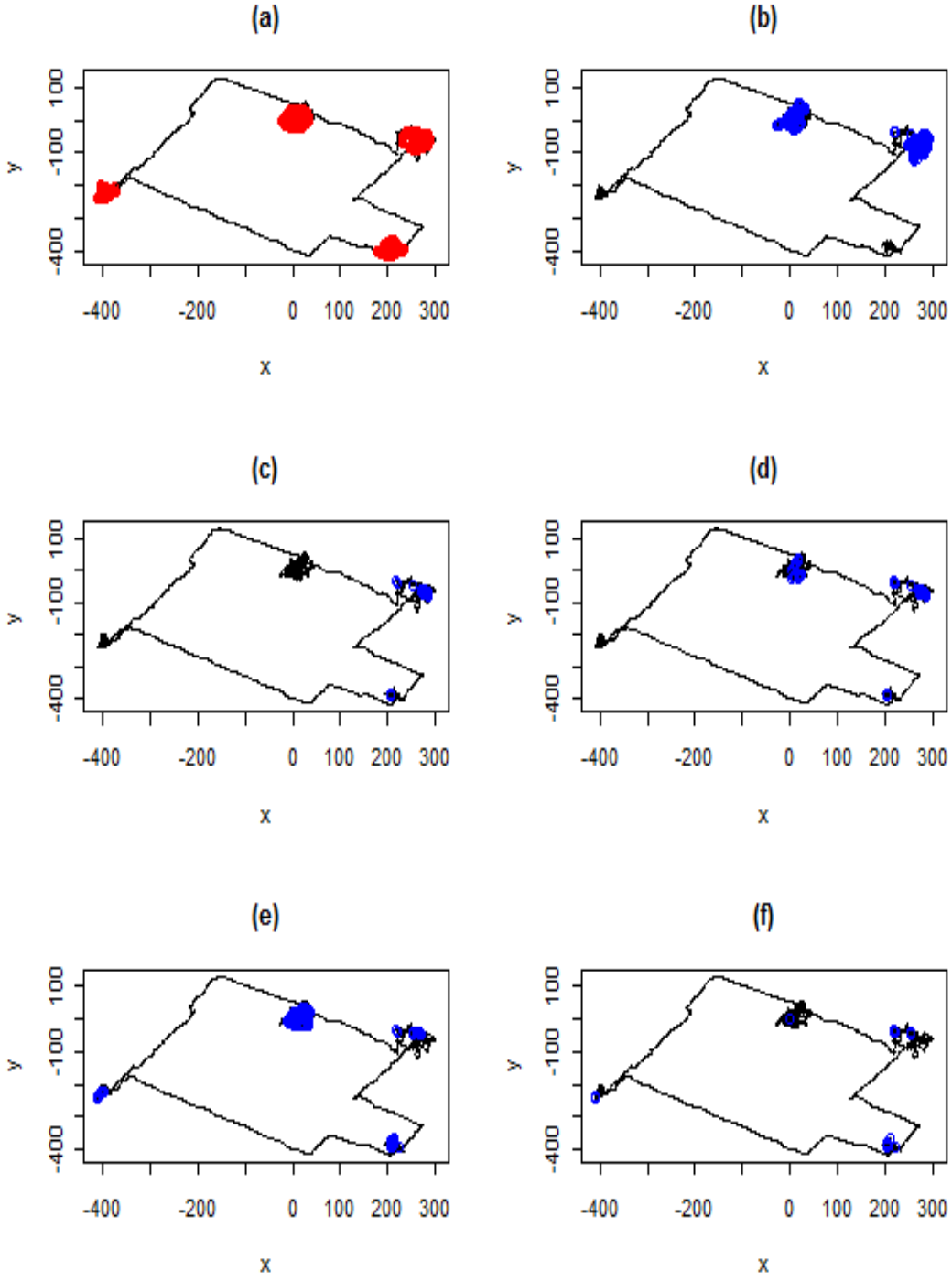

Figure 6.2: Plots for subject 1, time period 1, day 2 where the unfiltered series is in black, the identified clusters are in red and the identified noisy windows are in blue:
(a) Unfiltered time series with identified clusters,
(b) Average amplitude of acceleration method with $\kappa = 1.25$,
(c) Standard deviation of distance method with $\kappa = 2$,
(d) Standard deviation of the amplitude of acceleration method with $\kappa = 2$,
(e) Ratio of standard deviation to mean distance method with $\kappa = 1.25$ and,
(f) Ratio of standard deviation to mean amplitude of acceleration method with $\kappa = 2$.

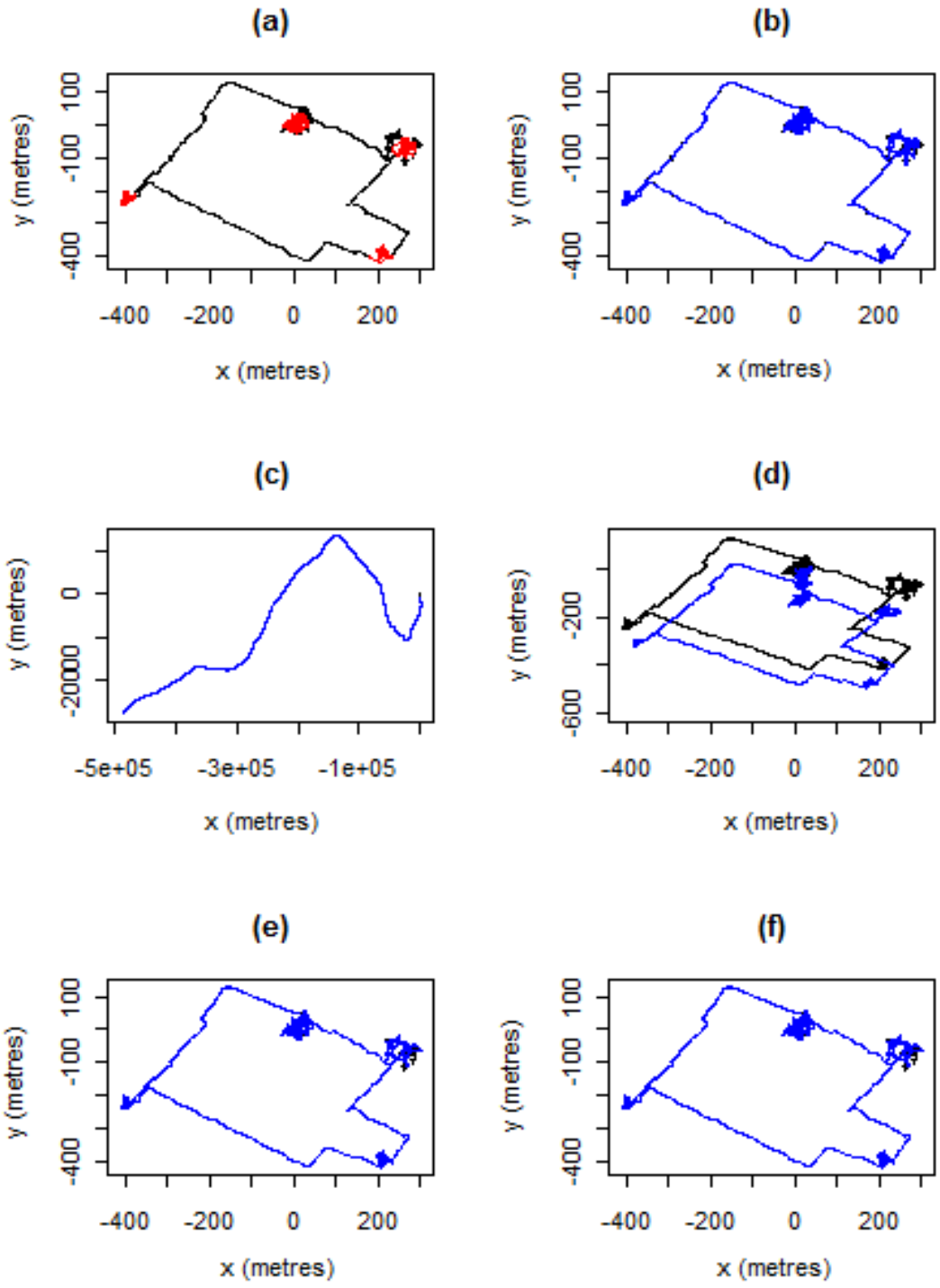

Figure 6.3: Plots for subject 1, time period 1, day 2 where the unfiltered series is in black, the identified clusters are in red and the filtered series are in blue:
(a) Unfiltered time series with identified clusters,
(b) Moving average with 21 points,
(c) Eliminating high accelerations with $\eta = 1.25$,
(d) Eliminating high velocities with $\eta = 2$,
(e) Trimmed means with 59 points and trim of 10% on each side and,
(f) Multilevel trimmed means with 59 points, $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$ and trimming parameters $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$.

locities with $\eta = 2$, (e) trimmed means with 59 points and trim of 10% on each side, and (f) multilevel trimmed means with 59 points, $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$ and trimming parameters $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.10, 0.15, 0.20)$ as a blue line.

From Figure 6.3 (c), it is clear that the shape of the time series of location points has been lost due to small errors accumulating throughout the integration process. Therefore, filtering out the high accelerations and setting them to zero is obviously not working well for filtering our location data. Similarly, from Figure 6.3 (d) it can be seen that although this method does not distort the shape of the time series, it does shift it, which is also not desirable. Figures 6.3 (b), (e) and (f) all look reasonable for the filtered series. This is due to the fact that this particular time series does not have a lot of large noise in it. Considering single large outliers in the location data can strongly influence the moving average, this method will also not be further considered.

Example 6.2. *Consider the time series of location points from subject 2, time period 1, day 2. Figure 6.4 (a) displays the unfiltered time series as a thin black line with the location points identified as part of a cluster in red. The individual has 2 identified clusters for this particular day and appears to have gone from one location to the other and returned using a similar route. There looks to be potential for very noisy windows, as there are some data points near the identified cluster at the bottom right hand corner that jut out to a peak and go back to the cluster.*

Figures 6.4 (b)-(f) display the unfiltered time series as a thin black line with the points in the windows identified as being very noisy by the average amplitude of acceleration method with $\kappa = 1.25$, standard deviation of distance method with $\kappa = 2$, standard deviation of the amplitude of acceleration method with $\kappa = 2$, ratio of standard deviation to mean distance method with $\kappa = 1.25$ and the ratio of standard deviation to mean amplitude of acceleration method with $\kappa = 2$, respectively, in blue. It can be seen that the methods with these particular κ values identify different win-

dows as being very noisy. Furthermore, they identify a different number of windows as being very noisy. However, it may be noted that most of the flagged windows occur at the identified cluster, which is consistent with the fact that we believe much of the noise in the series arises near the identified clusters.

It is clear that the various methods are identifying different windows of time with these given κ values. In Figure 6.4 (b), (d) and (f), the identified problematic windows occur near the clusters whereas in Figures 6.4 (c) and (e) numerous windows throughout the series are identified. As there are no obvious noise issues along the path between the two clusters, it appears the standard deviation of distance and ratio of standard deviation to mean distance methods are not performing as desired.

Figure 6.5 displays the filtered time series using the multilevel trimmed means method. This method has taken out the large spikes in location from the identified clusters. The bottom right corner now has a smoother appearance to it, which is to be expected and desired after filtering.

Example 6.3. *In this example, we will examine the time series of location points from subject 12, time period 1, day 7. Figure 6.6 (a) displays the unfiltered time series as a thin black line with the points in the identified cluster in red. It looks as if the individual remained at home throughout the day. However, it is very clear that there is an issue with noise in this data set. There are several data points that stray quite far from the centre of the identified cluster, as well as a large jump in the data from the (0,0) position to where the cluster is located.*

Figures 6.6 (b)-(f) display the unfiltered time series as a thin black line with the points in the windows identified as being noisy by the average amplitude of acceleration method with $\kappa = 1.25$, standard deviation of distance method with $\kappa = 2$, standard deviation of the amplitude of acceleration method with $\kappa = 2$, ratio of standard deviation to mean distance method with $\kappa = 1.25$ and ratio of standard deviation to mean

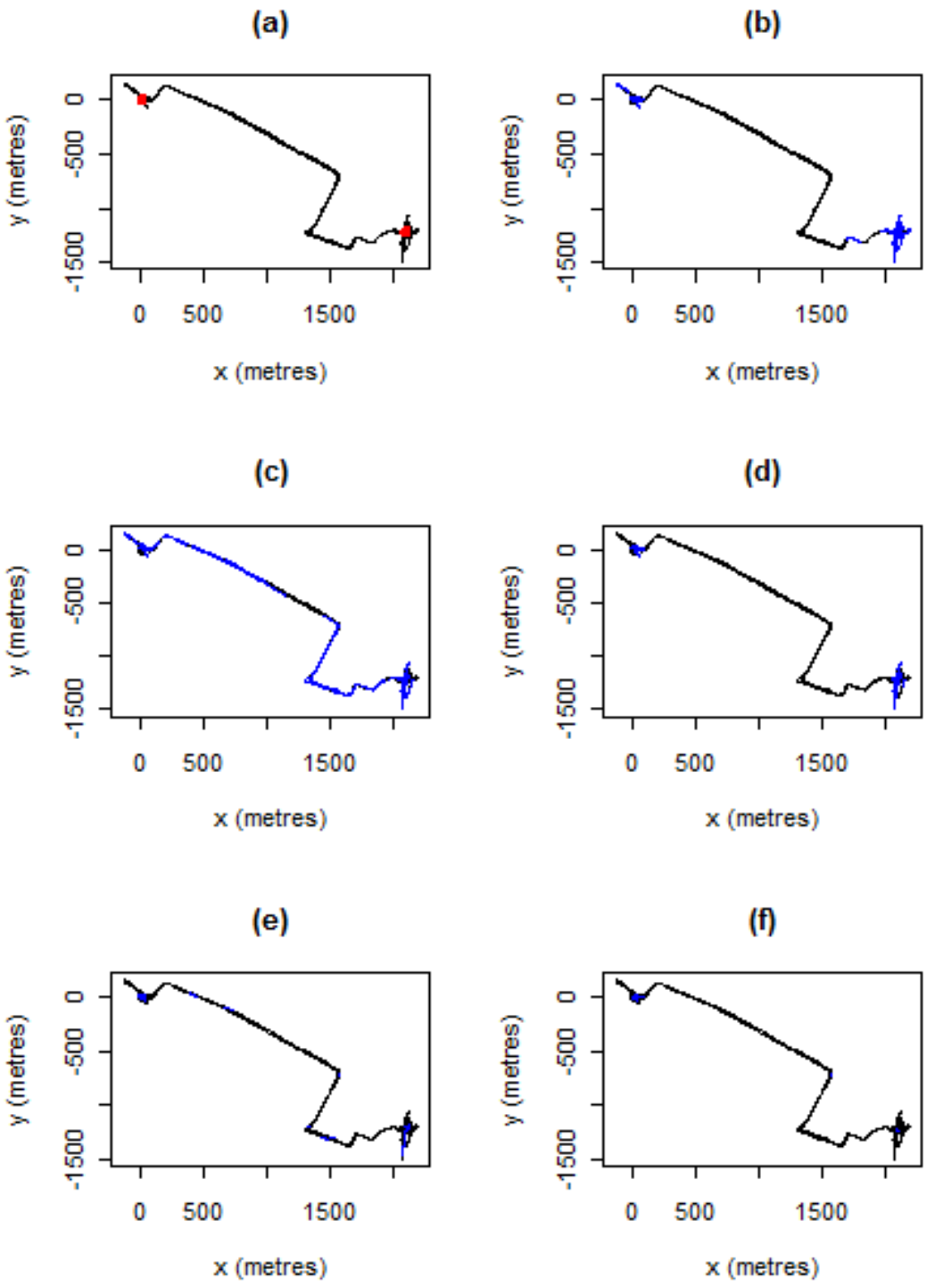

Figure 6.4: Plots for subject 2, time period 1, day 2 where the unfiltered series is in black, the identified clusters are in red and the identified noisy windows are in blue:
(a) Unfiltered time series with identified clusters,
(b) Average amplitude of acceleration method with $\kappa = 1.25$,
(c) Standard deviation of distance method with $\kappa = 2$,
(d) Standard deviation of the amplitude of acceleration method with $\kappa = 2$,
(e) Ratio of standard deviation to mean distance method with $\kappa = 1.25$ and,
(f) Ratio of standard deviation to mean amplitude of acceleration method with $\kappa = 2$.

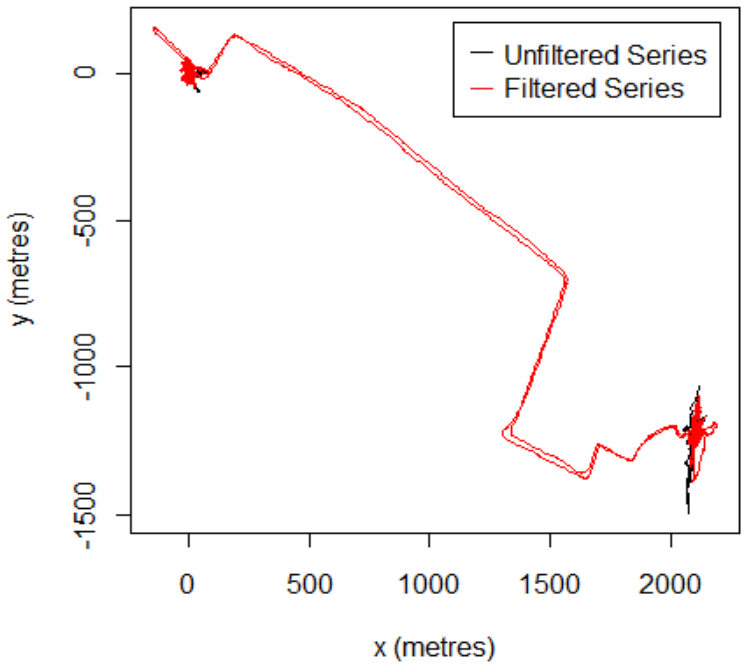

Figure 6.5: Plot of filtered times series for subject 2, time period 1, day 2 using the multilevel trimmed means method with 4 levels, $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$, and trimming parameters $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$.

amplitude of acceleration method with $\kappa = 2$, respectively, in blue. It can be seen that the methods with these particular κ values identify different windows as being noisy, as well as a different number of windows. The flagged windows occur at an identified cluster, which is consistent with the fact that we believe much of the noise in the series arises near clusters. From these plots, it is clear that much of this time series is being flagged as very noisy, which is to be expected based on how many points seem to be straying from the centre location.

Figure 6.7 displays the filtered data set using the multilevel trimmed means procedure with $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$. These parameter values were chosen as heuristically appear to give the smoothest series, which can be seen in Figure 6.8. The multilevel trimmed means filter with these parameters, presented in Figure 6.8 (a), was able to remove the first few points of the series located around $(0, 0)$, the large spike of data going out to $x = 200$ as well as the spike near $(-200, 100)$. The resulting series is not as smooth with the other parameter values and the large spike near $(200, 250)$ was not removed in Figure 6.8 (b) and (c), nor were the points at the beginning of the time series. However, there are still some issues that this multilevel trimmed means method was unable to deal with, such as the large jump from approximately the (0,0) location to the cluster. However, since this method deals with the majority of the large noise issues that arise, we will continue to use this method as our filtering technique.

6.5 Distance measurements

Various methods to identify where large noise is located as well as how to filter time series of location points have been discussed in the previous sections. Now that an identification method, average amplitude of acceleration, and filtering technique,

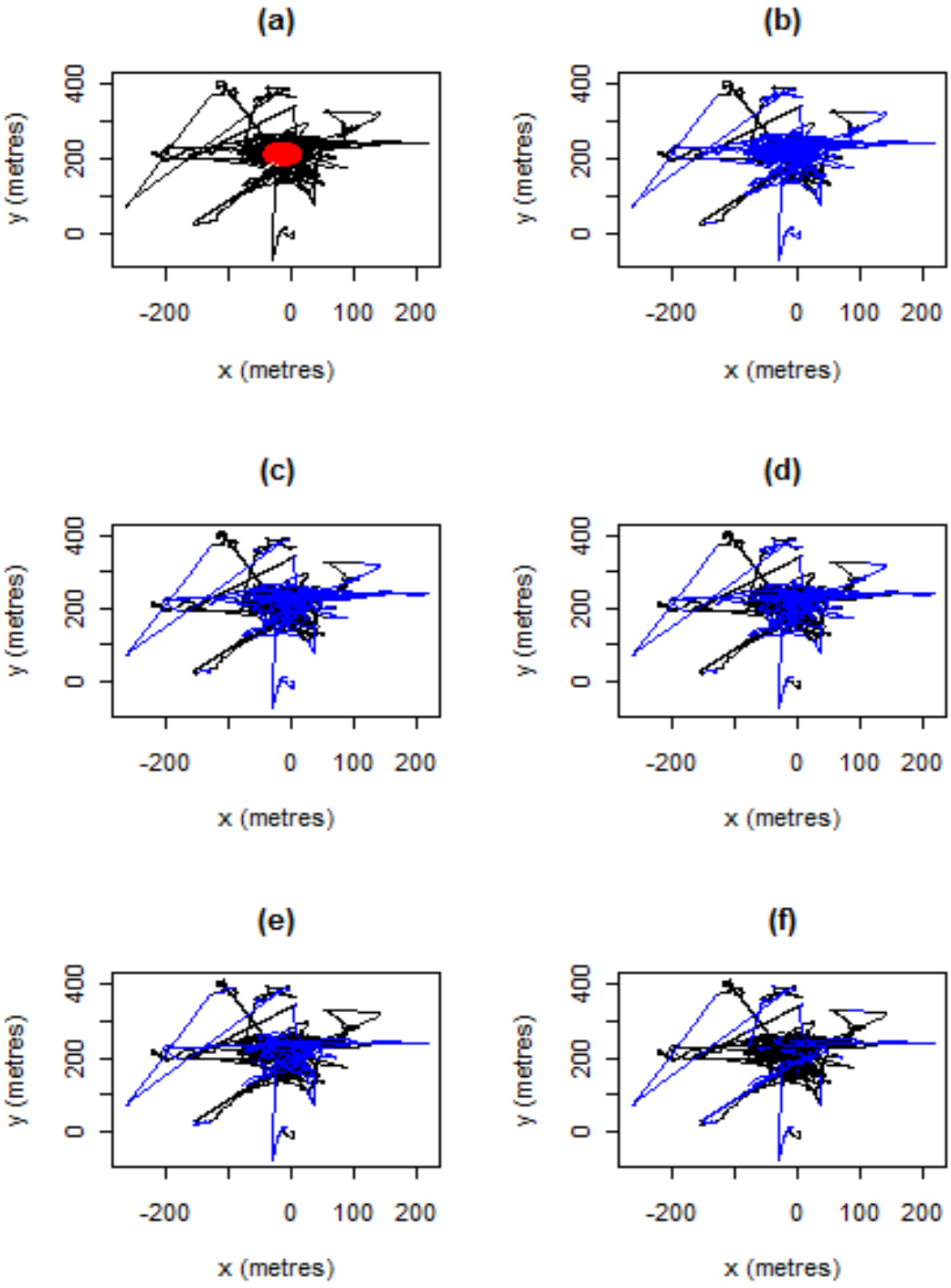

Figure 6.6: Plots for subject 12, time period 1, day 7 where the unfiltered series is in black, the identified cluster is in red and the filtered series are in blue:
(a) Unfiltered time series with identified clusters,
(b) Average amplitude of acceleration method with $\kappa = 1.25$,
(c) Standard deviation of distance method with $\kappa = 2$,
(d) Standard deviation of the amplitude of acceleration method with $\kappa = 2$,
(e) Ratio of standard deviation to mean distance method with $\kappa = 1.25$ and,
(f) Ratio of standard deviation to mean amplitude of acceleration method with $\kappa = 2$.

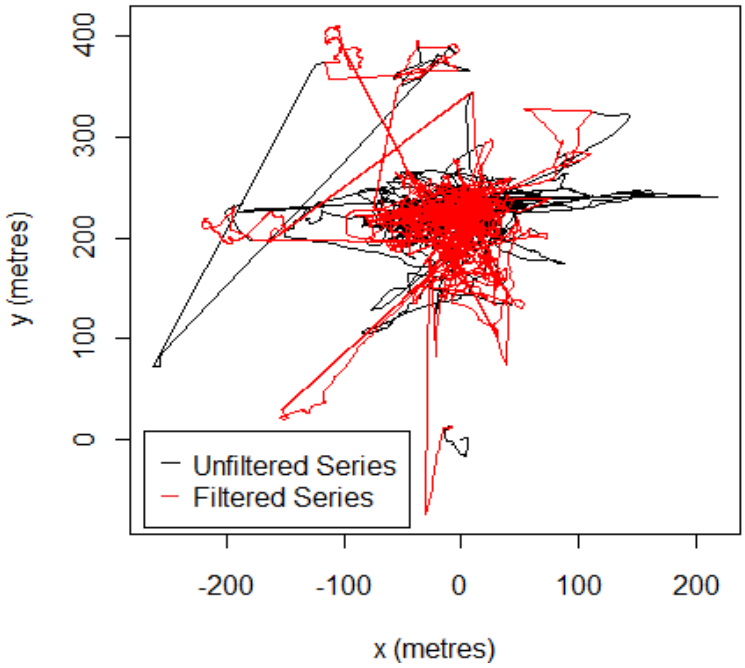

Figure 6.7: Plot of filtered times series for subject 12, time period 1, day 7 using the multilevel trimmed means method with 4 levels, $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$, and trimming parameters $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$.

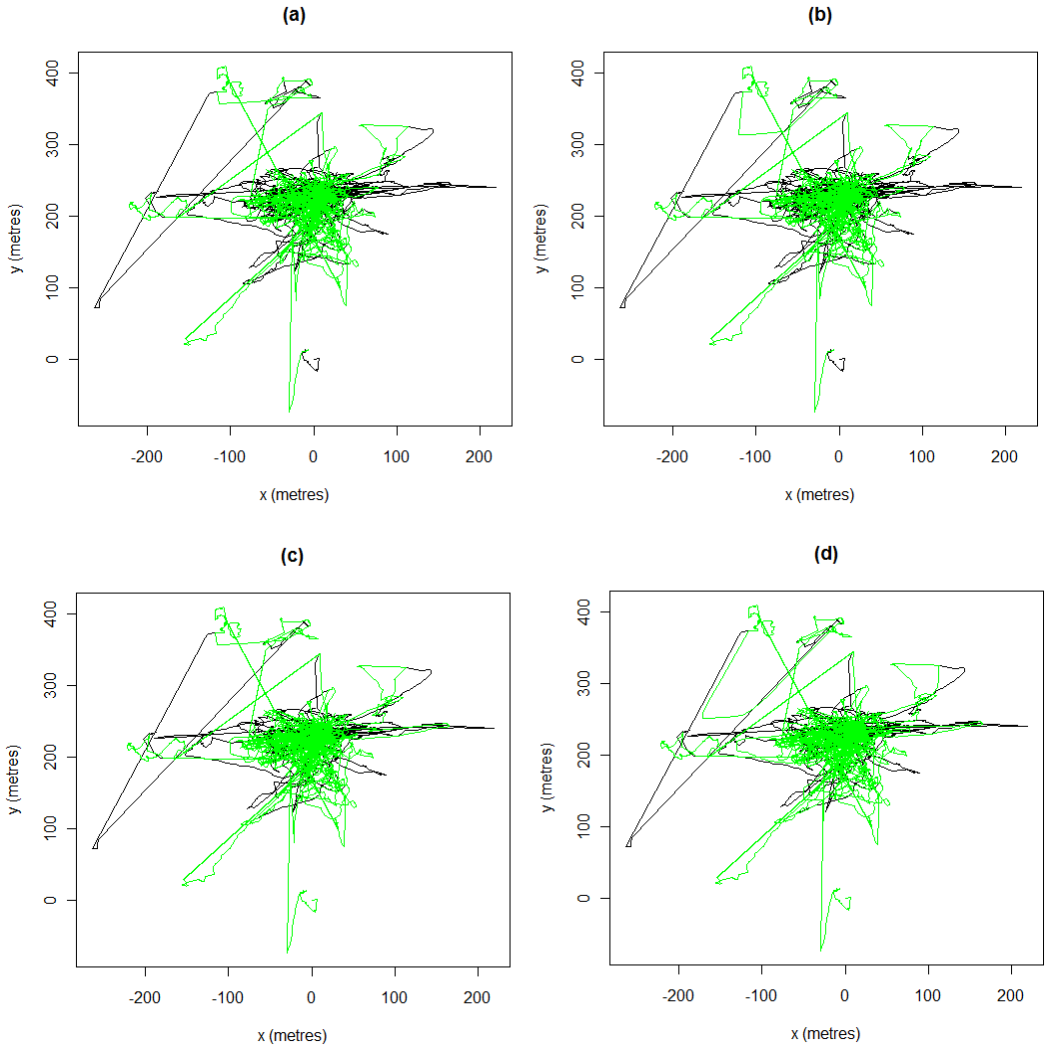

Figure 6.8: Plots for subject 1, time period 1, day 2 with the unfiltered series in black and the filtered series resulting from the mutlilevel trimmed means with the following noise cut-off values and smoothing parameters in green:
(a) $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2.0)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$,
(b) $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2.0)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.01, 0.02, 0.03, 0.04)$,
(c) $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (2.0, 2.25, 2.5, 2.75)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.1, 0.15, 0.2)$,
(d) $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (2.0, 2.25, 2.5, 2.75)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.1, 0.2, 0.3, 0.4)$.

multilevel trimmed means, have been chosen as appropriate methods for time series of location points, we will now recalculate the distance travelled by the individuals. As the series are now "clean", the resulting distances should be more accurate and approaching the true distance travelled.

The distances of the filtered series are calculated as follows:

- Compute the Euclidean distance between consecutive time points, denoted by d_{f_t}. The value d_{f_t} is calculated by $d_{f_t} = \sqrt{(\hat{x}_{t+1} - \hat{x}_t)^2 + (\hat{y}_{t+1} - \hat{y}_t)^2}$ for $t = 1, 2, \ldots, n-1$ where $\hat{x}_t$ and $\hat{y}_t$ are the x and y coordinates of the filtered series, respectively.

- Compute the total distance travelled throughout the day by summing the values of d_{f_t} for $t = 1, 2, \ldots, n-1$, and denote this by d_{f_T}. Hence, $d_{f_T} = \sum_{t=1}^{n-1} d_{f_t}$ is the total distance the individual travelled in metres.

- Divide d_{f_T} by 1000 to convert the distance to kilometres.

The updated distance results are based on identifying the large noise using the average amplitude of acceleration method and filtering the time series using the multilevel trimmed means method with $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (1.25, 1.5, 1.75, 2)$, $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.05, 0.10, 0.15, 0.20)$ and a window length of $l(w) = 30$ seconds. These values were chosen as they heuristically seem to give the most desirable and smoothest series, as in in Figure 6.8. These results are presented in Appendix C. The distances obtained from the filtered series are the same or lower than that of the unfiltered series. We will consider the distances computed from the filtered series to be an upper bound on the true total distance travelled, as the filtered series will still contain some noise due to our filtering method being unable to fix all potential noise problems.

Now, let us take a look at two examples of how this recalculated distance compares to the originally calculated distance on the unfiltered data set.

Example 6.4. *Consider the city block example that we have seen many times throughout this thesis (subject 1, time period 1, day 2). The series had very few windows identified as being very noisy and therefore, one would expect the recalculated distance to be quite similar to the originally calculated total distance travelled. From the results presented in Appendix C, the unfiltered total distance was 10.2 kilometres and the filtered total distance was reduced to 9.8 kilometres, which is consistent with the expectation that the distance measurement would not change too much. Therefore, the distance measurement was reduced by 400 metres, which is less than 4% of the original distance.*

Example 6.5. *Consider the at home example presented in Example 6.3. In this series, there were several windows identified as having large amounts of noise, meaning one would expect that the recalculated distance is substantially lower than the unfiltered total distance. From the results presented in Appendix C, the unfiltered total distance was 26 kilometres and the filtered total distance was calculated to be 17.6 kilometres. Hence, the distance travelled was reduced by 8.4 kilometres, which is almost a third of the original calculated total distance. However, this still appears to be quite a high value for the distance travelled for an individual who appears to have remained near the home location for the day. This may be due to the fact that there is still some remaining noise and this time series has over 13 hours of recordings making it such that even the little distances travelled around the house add up.*

Figure 6.9 displays a boxplot of the total distance travelled based on the filtered time series of location points. It is clear that the total distance ranges greatly. For instance, participant 15 has total distances ranging from 0.6 kilometres to 598.3 kilometres. Figure 6.10 (a) displays the graph of the distances from the unfiltered series versus the filtered series with a 45 degree line. There is naturally a strong linear relationship between two distances calculated. The very large distances did not change

very much after filtering the series, whereas some of the shorter distances were greatly reduced. Figure 6.10 (b) is the same plot zoomed in on the shorter distances to display the series in which the filtering significantly effected the total distance.

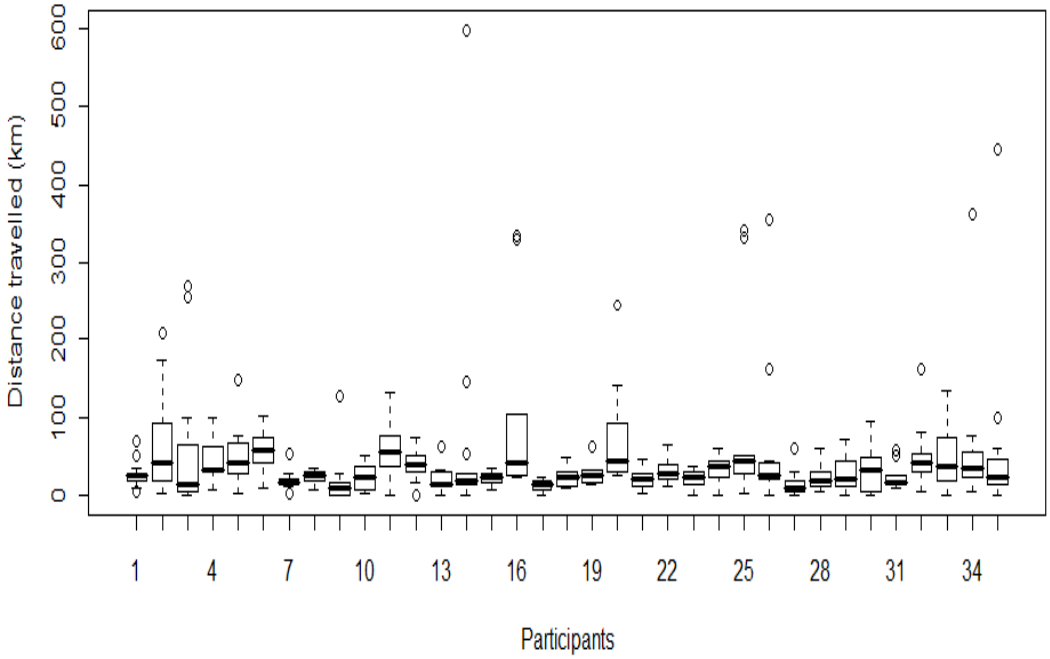

Figure 6.9: Boxplot of the distances travelled using the filtered series.

A lower bound will also be computed for the true distance travelled. One possible lower bound to consider is computed as the sum of the Euclidean distances between the cluster centres. For instance, consider the case when an individual has four identified clusters with cluster centres (x_{c_1}, y_{c_1}), (x_{c_2}, y_{c_2}), (x_{c_3}, y_{c_3}) and (x_{c_4}, y_{c_4}). Further consider that the individual's route that day had him/her stop at cluster 1, 2, 3, 4, and then return to cluster 1 again. The lower bound on the distance would be computed as follows:

$$\begin{aligned} d_L = &\sqrt{(x_{c_2} - x_{c_1})^2 + (y_{c_2} - y_{c_1})^2} \; + \; \sqrt{(x_{c_3} - x_{c_2})^2 + (y_{c_3} - y_{c_2})^2} + \\ &\sqrt{(x_{c_4} - x_{c_3})^2 + (y_{c_4} - y_{c_3})^2} \; + \; \sqrt{(x_{c_1} - x_{c_4})^2 + (y_{c_1} - y_{c_4})^2}. \end{aligned}$$

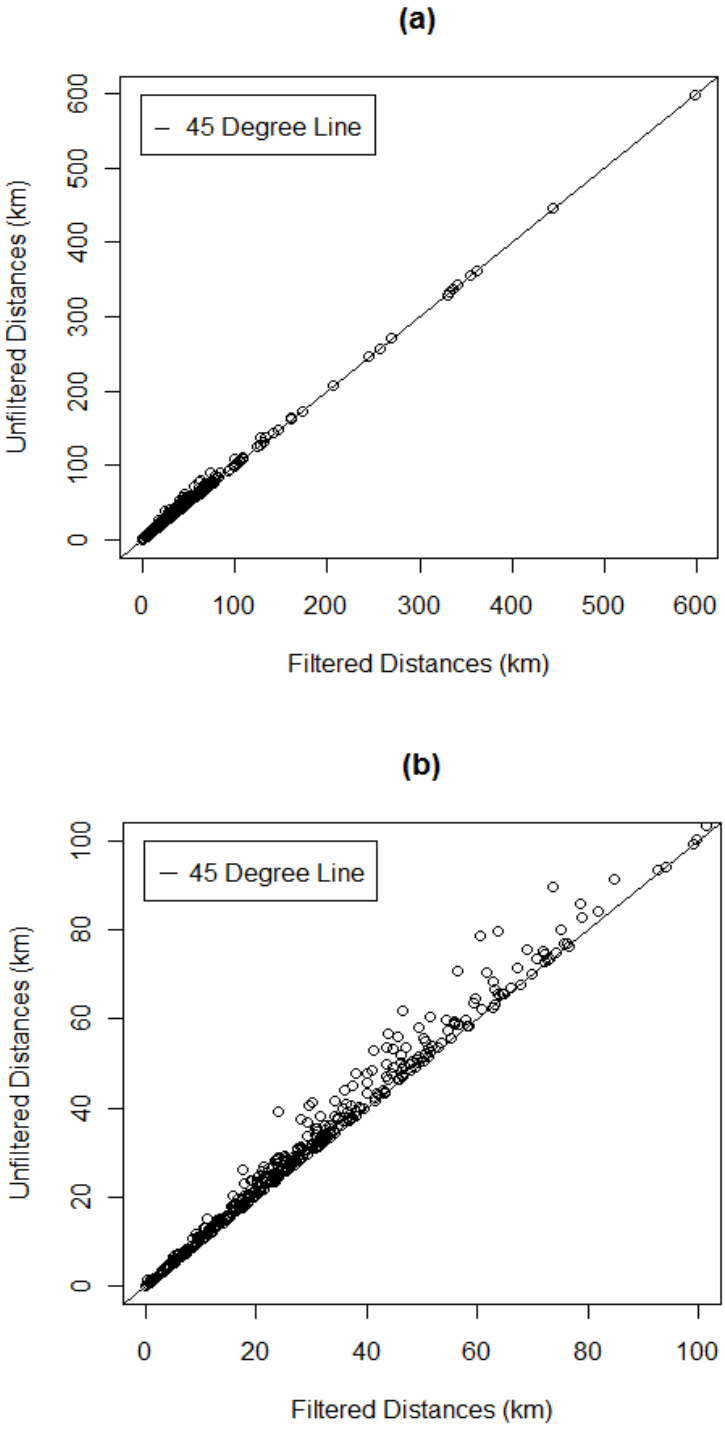

Figure 6.10: Total distance calculated from unfiltered time series vs. total distance calculated from filtered time series with 45 degree line representing equality:
(a) All data points, (b) Zoomed-in on shorter distances.

However, this is a very crude lower bound, and a more appropriate bound may be computed. The lower bound that will be considered in this thesis is calculated as follows:

1. Set all points in cluster i to (x_{c_i}, y_{c_i}), the central value of the cluster. Leave all points not identified as being in a cluster where they are. Denote the x and y values of the resulting time series by x_{cent} and y_{cent}, respectively.

2. Compute the Euclidean distance between consecutive time points of the centred series. i.e. $d_{cent_t} = \sqrt{(x_{cent_{t+1}} - x_{cent_t})^2 + (y_{cent_{t+1}} - y_{cent_t})^2}$ for $t = 1, 2, \ldots, n-1$, where n is the length of the interpolated and filtered series.

3. Sum the d_{cent_t} values and divide by 1000 to convert the distance to kilometres. i.e. $d_{cent_T} = \sum_{t=1}^{n-1} d_{cent_t}$

This lower bound was chosen as an appropriate bound due to large noise occurring in the clusters. By setting the distance travelled among the points in the identified clusters to zero, the distance added from the large noise is eliminated. Furthermore, the true distance travelled in a single cluster is expected to be small meaning that by setting it to zero, there is little true travelled distance lost.

Figure 6.11 (a) displays the graph of the distances from the filtered series in red along with the lower bound on the distances in blue. This allows one to see the difference between the distance of the filtered series and the computed lower bound. Figure 6.11 (b) displays the graph of the distances from the filtered series vs. the lower bound on distance with a 45 degree line. There is naturally to be a strong linear relationship between two distances calculated. The very large distances for the filtered series are quite similar to the corresponding lower bounds, whereas some of the shorter distances were greatly reduced.

The lower bounds on the total distance are presented in Appendix C. Note that

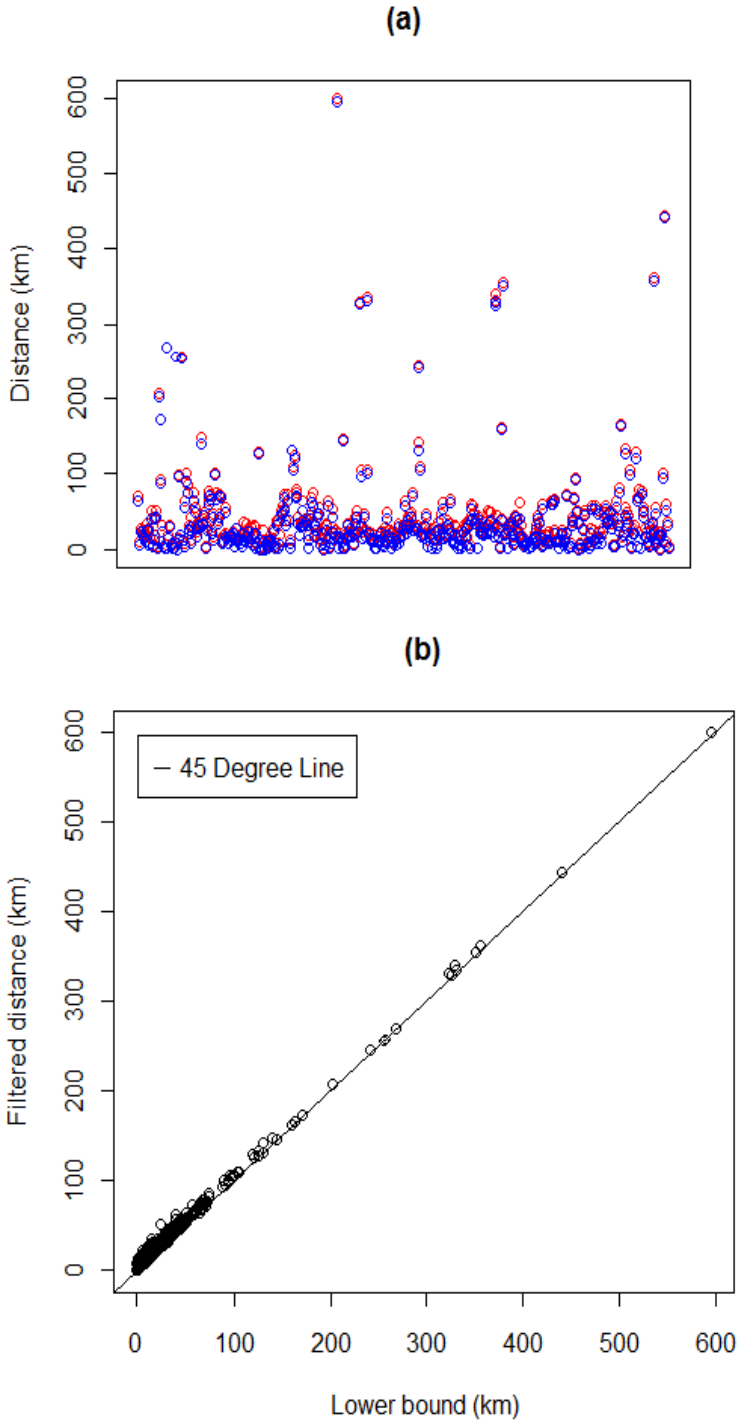

Figure 6.11: Plots displaying the filtered distance in comparison to the lower bound: (a) Total distance of filtered series in red and lower bound on total distance in blue, (b) Total distance of filtered series vs. lower bound on total distance with 45 degree line representing equality.

if the interval (lower bound, upper bound) has a small range, we have a good idea of the true distance. However, if the range of the interval (lower bound, upper bound) is large, we are uncertain as to where within the interval the true distance falls.

Chapter 7

Discussion and Conclusions

The concept of community mobility is important in the medical field, especially to those working with the elderly or patients affected by various neurological diseases. Various aspects of mobility are investigated in this thesis and clustering, large noise identification and filtering techniques are discussed.

Previously developed clustering techniques, k-means and trimmed k-means, are presented and some basic properties are discussed. If the number of clusters is known and appropriate trimming level is chosen, the trimmed k-means method identifies the location of the clusters very well. However, in practical applications of location data from an individual's daily movement, the number of clusters/hotspots are unknown. By using traditional clustering techniques, the time dependencies are ignored and therefore it is unknown whether the location is truly a cluster or if the individual just happened to pass by that particular location multiple times that day, or how many times the individual visited the same location. Since the time dependencies are known, it is best to incorporate them into the identification of the location of each cluster.

Boissy et al. (2012) proposed a method of identifying the locations of hotspots

in GPS time series. In this clustering method a grid was superimposed over the area the individual occupied. The grid was set up such that each cell block corresponds to a 0.005 degree change in latitude and longitude. A cell was flagged as a cluster if the individual remained in the grid's cell block for a minimum of five minutes. Once all of the cell blocks of the grid had been classified as a cluster or not, a second step of clustering occurs in which adjacent flagged cell blocks are joined. This method has a few downfalls including but not limited to the fact that it may miss a cluster if it hovers at the borders of the grid due to some points being located in one cell block and some in the adjacent cell block.

In this thesis we have developed a new effective clustering method that accounts for the time dependencies and does not require the number of clusters to be previously known. Furthermore, it does not impose a spatial grid requirement and instead uses a circle around the central point of the window making it such that clusters will not be missed but rather they will be identified using scrolling time windows. Sensitivity analysis was performed on this proposed method and indicated that the method is not sensitive to the choice of the parameters (trim level, window size, radius of circle).

It should be noted that although some GPS time series have clusters that can be easily identified by visual inspection, this is not the case for all time series and therefore an automated identification method is required. Furthermore, by using only visual inspection there is potential to identify more or fewer clusters than are truly in the data. For example, if an individual passes through the same location numerous times a day but does not stop there, by visual inspection only, it may appear as if there is a cluster located at that location. On the other hand, if an individual remains at the same location for an extended period of time and the GPS signal is very accurate it may be impossible to tell by visual inspection if there are thousands of points located in precisely the same location or not.

In practice, large noise arises in GPS data sets, and can greatly influence the results. Boissy et al. (2012) discuss a technique commonly used to filter GPS time series. In order to gain a smoother time series of location points, one can remove all data points that have an accuracy below a given cut-off value. Once the points have been removed, linear interpolation may be implemented to complete the smoothed time series. Various methods to identify large noise in the time series were explored in this thesis and the average amplitude of acceleration was suggested as the best identifier. To achieve more accurate results in our analysis, the windows of time identified as being very noisy were filtered to smooth the series. The adaptive multi-level trimmed means filtering method, which depends on the amount of noise within the series, is shown to be very effective in the analysis. These methods work well for the GPS time series as they do not require the precision of the data to be known, which is beneficial as not all GPS devices store the accuracy of each point.

Results for various measures of mobility are presented in this thesis. The total distance travelled and a lower bound on the distance, the number of identified clusters, as well as the proportion of time spent in the identified clusters and the area of a robust and classical 95% ellipse and a minimum spanning ellipse, better known as lifespace, are provided for all subjects included in the mobility study. The proposed measures for the lower bound on distance and the construction of the minimum spanning ellipse for the time series involve setting all points in each identified cluster to each cluster's central point. Due to the fact that the large noise typically occurs in the clusters, these methods are robust in the sense that they eliminate the wandering around inside the clusters. The lower bound on distance is a compromise between computing the distance with all the data and the sum of the Euclidean distance between the centre points of consecutive clusters in time. As described by Boissy et al. (2011), the classical 95% ellipse based on the mean and covariance structure is the el-

lipse construction used in practice currently. However, the minimum spanning ellipse outperforms the classical ellipse due to the fact that it does not have an underlying normality assumption, does not cut out portions of the data where the individual travelled and does not extend well beyond any recorded location points. A measure of accuracy of the data is the proportion of recorded time, as it indicates the amount of necessary linear interpolation. It can be noted that improved measurements lead to a more accurate characterization of an individual's mobility.

Additionally, further studies can be done to analyze the total number of stops an individual makes throughout the day rather than the number of distinct stops. A new method for computing the 95% ellipse for the data set that includes all of the points in an identified hotspot and has minimal area where points are not located may be investigated. With more information regarding the status of the individuals, comparisons could be made between groups (male versus female, Parkinson's patients versus controls, different age categories, etc.).

Bibliography

[1] Afifi, A., Clark, V. A., and May, S. (2004). *Computer-Aided Multivariate Analysis*, Chapman & Hall/CRC, Boca Raton.

[2] Arden, B. and Astill, K. (1970). *Numerical Algorithms: Origins and Applications*, Addisson-Wesley Publishing Company, Massachusetts.

[3] Baker, P.S., Bodner, E. V., and Allman, R. M. (2003). Measuring life-space mobility in community-dwelling older adults. *Journal of the American Geriatrics Society*, 51: 1610-1614.

[4] Boissy, P., Brière, S., Gringras-Hill, C., Blamoutier, M., Cabana, F., Duval, C. and the EMAP group (2012). GPS and intertial sensor data fusion for ecological assessment of lifespace and mobility constriction in aging and disease. *Preprint*.

[5] Boissy, P., Brière, S., Hamel, M., Jog, M., Speechley, M., Karelis, A., Frank, J., Vincent, C., Edwards, R., Duval, C. and the EMAP group (2011). Wireless inertial measurement unit with GPS (WIMU-GPS) - Wearable monitoring platform for ecological assessment of lifespace and mobility in aging and disease. *Annual International Conference of the Institute of Electrical and Electronics Engineers (IEEE) Engineering in Medicine and Biology Society (EMBS)*, 33: 5815-5819.

[6] Burr, I. W. (1979). *Elementary Statistical Quality Control*, Marcel Dekker, Inc., New York.

[7] Cuesta-Albertos, J. A. and Matràn, C. (1997). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probability Theory and related Fields*, 78: 523-534.

[8] Dillon, W. R. and Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*, John Wiley & Sons, Inc., New York.

[9] Freire dr Mello, L. and Marandola Jr, E.J. (2005, July). *Life spaces, mobility and the metropolis: dialoguing with geography.* Paper presented at XXV International Union for the Scientific Study of Population (IUSSP) International Population Conference, Tours, France.

[10] Garcia-Escudero, L. A. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447): 956-969.

[11] Gnanadesikan, R. (1977) *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons, Inc., New York.

[12] Hartigan, J. A. (1975). *Clustering Algorithms*, John Wiley & Sons, Inc., New York.

[13] Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28: 100-108.

[14] Jamieson, A. (1982). *Introduction to Quality Control*, Reston Publishing Company, Inc., Virginia.

[15] Kaplan, E.D., Hegarty, C.J. (2006). *Understanding GPS Principles and Applications: Second Edition*, Artech House, Inc., Massachusetts.

[16] Masumoto, Y., Pioneer Electronic Corporation (1993) *Global Positioning System*, U.S. Patent No. 5210540.

[17] Montgomery, D. C. (1985). *Introduction to Statistical Quality Control*, John Wiley & Sons, Inc., New York.

[18] Parkinson, B.W. and Spilker, J.J.Jr. (1996). *Global Positioning System: Theory and Applications Volume I*, American Institute of Aeronautics and Astronautics, Inc., Virginia.

[19] Parkinson, B.W. and Spilker, J.J.Jr. (1996). *Global Positioning System: Theory and Applications Volume II*, American Institute of Aeronautics and Astronautics, Inc., Virginia.

[20] Pham, D.T., Dimov, S.S. and Nguygen, C.D. (2005). Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers*, 1: 103-119.

[21] Pyzdek, T. (1989). *What Every Engineer Should Know About Quality Control*, Marcel Dekker, Inc., New York.

[22] Radi, H.R., Vivian, L.W.B., Zainudin, M.N.S. and Ismail, M.M. (2012). FPGA-Based Global Positioning System. *International Journal of Electrical and Computer Sciences*, 12: 11-14.

[23] Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41: 212-223.

[24] Shumway, R.H. and Stoffer, D.S. (2006). *Time Series Analysis and Its Applications: With R Examples*, Springer, New York.

[25] Spilker, J.J.Jr. (1978). GPS signal and performance characteristics. *Navigation*, 25: 121-146.

[26] Titterington, D.M. (1978). Estimation of Correlation Coefficients by Ellipsoidal Trimming. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27: 227-234.

[27] Rick. (08/10/2004). The Math Forum: Ask Dr. Math. April 25, 2012, http://mathforum.org/library/drmath/view/51833.html.

[28] Tsui, J.B.-Y. (2005). *Fundamentals of Global Positioning System Receivers: A Software Approach*, John Wiley & Sons, Inc., New Jersey.

[29] Wells, D.e., Beck, N., Delikaraoglou, D., Kleusberg, A., Krakiwsky, E.J., Lachapelle, G., Langey, R.B., Nakiboglu, M., Schwarz, K.P., Tranquilla, J.M., Vanicek,P. (1999). Guide to GPS positioning. *Canadian GPS Associates*, New Brunswick.

Appendix A

Distance travelled (km) (D), length of time series (T), proportion of recorded time (P)

Table A.1: Distance, Time and Proportion

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
1	D	73.5	10.2	30.7	29.0	20.3	27.7	25.8	NA
Period 1	T	10:14:32	4:53:57	10:56:30	8:36:19	8:6:31	10:39:15	8:21:11	NA
	P	0.473	0.768	0.573	0.750	0.380	0.824	0.918	NA
1	D	38.1	30.9	23.4	17.8	55.5	5.0	NA	NA
Period 2	T	13:47:9	7:58:41	12:13:31	11:58:43	12:14:46	1:26:48	NA	NA
	P	0.619	0.715	0.737	0.522	0.717	0.888	NA	NA
2	D	43.4	21.6	53.8	2.5	NA	NA	NA	NA
Period 1	T	13:56:05	24:00:00	22:07:59	4:28:15	NA	NA	NA	NA
	P	0.778	0.903	0.819	0.983	NA	NA	NA	NA
2	D	207.8	93.4	172.3	31.9	4.0	NA	NA	NA
Period 2	T	11:21:40	9:31:54	11:56:12	13:42:03	0:38:52	NA	NA	NA
	P	0.591	0.852	0.737	0.584	0.956	NA	NA	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
3	D	0.7	270.5	5.3	17.4	33.6	6.4	4.5	8.7
Period 1	T	1:00:41	10:59:33	8:53:45	9:20:32	10:14:06	14:32:00	10:30:15	7:28:48
	P	0.919	0.683	0.709	0.476	0.522	0.122	0.514	0.257
3	D	0.8	256.6	7.4	15.4	99.5	18.4	256.5	12.5
Period 2	T	1:29:15	8:22:27	8:13:20	7:46:25	13:31:29	13:47:42	8:40:43	6:28:48
	P	0.865	0.633	0.729	0.433	0.645	0.127	0.702	0.294
4	D	7.3	68.4	109.3	91.3	33.4	43.7	NA	NA
Period 1	T	6:27:39	13:51:03	15:48:51	14:10:26	14:07:27	16:02:30	NA	NA
	P	0.923	0.707	0.601	0.839	0.672	0.518	NA	NA
4	D	11.9	31.5	79.9	58.7	36.0	28.5	33.4	NA
Period 2	T	6:13:42	15:48:44	15:30:46	15:49:59	13:07:12	10:12:45	9:04:17	NA
	P	0.839	0.564	0.680	0.566	0.629	0.609	0.756	NA
5	D	31.4	148.2	39.9	53.3	57.9	4.1	NA	NA
Period 1	T	9:21:15	15:01:28	15:45:00	16:05:30	11:47:26	15:55:40	NA	NA
	P	0.770	0.862	0.747	0.683	0.656	0.271	NA	NA
5	D	29.6	77.1	63.1	75.2	28.5	17.4	NA	NA
Period 2	T	8:25:48	16:56:27	24:11:59	12:58:38	14:11:40	7:41:34	NA	NA
	P	0.860	0.708	0.387	0.727	0.658	0.988	NA	NA
6	D	46.4	103.4	74.9	37.8	47.2	NA	NA	NA
Period 1	T	13:11:31	23:07:32	14:36:08	12:36:16	10:42:32	NA	NA	NA
	P	0.768	0.842	0.656	0.635	0.730	NA	NA	NA
6	D	73.6	70.0	11.5	NA	NA	NA	NA	NA
Period 2	T	11:54:31	13:12:36	2:22:20	NA	NA	NA	NA	NA
	P	0.758	0.711	0.977	NA	NA	NA	NA	NA
7	D	20.4	26.7	57.4	17.8	17.1	20.6	23.8	NA
Period 1	T	9:48:47	7:34:19	13:45:41	5:46:23	7:10:42	10:40:38	7:59:10	NA
	P	0.470	0.497	0.664	0.356	0.769	0.291	0.412	NA
7	D	16.7	28.2	3.9	31.4	20.8	12.0	14.4	14.9
Period 2	T	5:59:10	8:38:13	4:27:07	9:06:45	7:46:16	8:01:30	6:15:45	5:26:05
	P	0.633	0.526	0.246	0.409	0.607	0.158	0.295	0.706
8	D	13.4	24.9	37.3	24.4	38.1	28.3	NA	NA
Period 1	T	9:33:29	15:12:18	9:18:01	11:38:12	12:01:02	8:29:04	NA	NA
	P	0.545	0.903	0.890	0.507	0.908	0.480	NA	NA
8	D	23.4	26.9	31.0	18.5	32.9	8.7	NA	NA
Period 2	T	8:29:04	7:00:55	15:00:57	14:33:47	11:32:14	5:30:41	NA	NA
	P	0.883	0.970	0.922	0.587	0.565	0.709	NA	NA
9	D	29.6	17.6	127.9	0.3	0.7	0.2	NA	NA
Period 1	T	12:13:56	23:27:07	16:15:25	15:19:07	24:00:00	5:48:57	NA	NA
	P	0.638	0.724	0.997	0.405	1.00	1.00	NA	NA
9	D	26.5	0.8	7.4	12.2	6.7	11.9	14.2	8.3
Period 2	T	11:00:57	19:42:53	21:08:31	13:10:33	12:21:53	12:16:47	4:35:31	1:33:37
	P	0.862	0.702	0.615	0.257	0.832	0.182	0.470	0.398

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
10	D	10.5	28.0	3.2	37.6	37.5	8.1	19.9	50.5
Period 1	T	8:45:08	15:10:41	3:58:31	11:58:31	11:07:14	13:20:16	10:49:58	8:07:38
	P	0.513	0.424	1.00	0.500	0.456	0.567	0.814	0.613
11	D	58.6	53.6	73.2	37.0	74.5	40.3	13.0	1.5
Period 1	T	13:32:35	23:40:07	13:42:32	12:23:16	12:16:16	15:20:27	16:49:20	0:58:46
	P	0.730	0.223	0.725	0.765	0.723	0.264	0.374	0.779
11	D	131.9	109.2	125.1	76.9	82.9	43.0	19.4	NA
Period 2	T	13:16:57	15:34:33	9:48:45	16:39:03	13:20:47	13:53:59	5:08:56	NA
	P	0.547	0.580	0.773	0.561	0.639	0.510	0.144	NA
12	D	44.0	79.8	40.5	47.8	43.4	48.5	26.0	1.4
Period 1	T	13:09:08	14:13:12	11:54:23	15:46:24	9:21:46	10:51:13	13:17:53	0:10:23
	P	0.551	0.785	0.524	0.380	0.369	0.683	0.487	0.995
12	D	49.9	71.5	89.7	36.0	19.5	35.5	53.5	59.6
Period 2	T	13:19:56	12:03:12	15:05:11	12:43:46	15:18:15	10:32:59	14:45:38	27:19:56
	P	0.467	0.642	0.701	0.153	0.128	0.582	0.352	0.503
13	D	11.7	15.2	28.0	0.2	NA	NA	NA	NA
Period 1	T	13:46:37	24:00:00	24:00:00	3:56:10	NA	NA	NA	NA
	P	0.764	0.884	0.806	1.00	NA	NA	NA	NA
13	D	15.2	34.8	32.2	65.7	3.1	15.7	NA	NA
Period 2	T	11:54:24	14:33:38	11:35:19	13:42:33	11:28:17	6:20:00	NA	NA
	P	0.738	0.687	0.890	0.793	0.862	0.808	NA	NA
14	D	28.7	54.5	22.3	26.5	597.9	0.6	NA	NA
Period 1	T	10:34:37	12:22:35	11:48:49	8:22:47	13:31:57	0:54:40	NA	NA
	P	0.875	0.697	0.603	0.807	0.752	0.943	NA	NA
14	D	15.1	20.2	14.8	3.1	147.3	5.7	20.3	NA
Period 2	T	6:29:37	9:56:09	9:23:55	9:47:10	8:14:14	6:19:21	6:50:59	NA
	P	0.302	0.598	0.564	0.587	0.601	0.805	0.702	NA
15	D	41.4	24.3	19.6	37.3	26.8	8.0	NA	NA
Period 1	T	10:04:33	9:40:31	6:08:32	9:44:00	6:43:04	6:59:39	NA	NA
	P	0.503	0.410	0.465	0.463	0.672	0.142	NA	NA
16	D	42.4	23.3	53.3	33.4	329.3	107.3	NA	NA
Period 1	T	11:40:36	4:21:13	8:00:04	13:56:24	13:05:01	11:41:22	NA	NA
	P	0.852	0.802	0.922	0.813	0.975	0.862	NA	NA
16	D	35.0	24.0	47.5	26.0	336.8	105.0	23.8	NA
Period 2	T	7:54:47	10:57:35	10:13:48	10:09:12	12:17:35	13:41:12	3:21:52	NA
	P	0.580	0.736	0.687	0.887	0.843	0.811	0.935	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
17	D	20.2	22.4	13.3	15.3	1.3	NA	NA	NA
Period 1	T	9:26:28	10:08:20	12:10:29	19:55:27	0:28:36	NA	NA	NA
	P	0.887	0.868	0.926	0.963	0.998	NA	NA	NA
17	D	0.5	12.7	18.6	9.0	27.4	NA	NA	NA
Period 2	T	4:15:49	4:47:09	7:32:38	3:18:05	7:56:14	NA	NA	NA
	P	0.036	0.882	0.639	0.974	0.677	NA	NA	NA
18	D	10.5	11.6	24.5	30.8	50.9	NA	NA	NA
Period 1	T	11:50:19	13:21:45	14:31:05	13:02:20	13:19:08	NA	NA	NA
	P	0.805	0.865	0.943	0.789	0.746	NA	NA	NA
18	D	26.6	26.1	23.3	14.3	21.9	7.9	NA	NA
Period 2	T	10:09:55	15:09:56	12:56:15	14:24:00	12:23:04	2:23:19	NA	NA
	P	0.794	0.751	0.822	0.858	0.825	0.974	NA	NA
19	D	14.2	21.3	14.2	35.5	14.2	NA	NA	NA
Period 1	T	12:41:47	11:57:18	12:09:49	13:52:52	18:39:28	NA	NA	NA
	P	0.658	0.754	0.752	0.866	0.882	NA	NA	NA
19	D	24.5	31.7	32.7	28.1	63.5	33.0	NA	NA
Period 2	T	12:57:11	11:37:28	12:35:55	13:26:52	14:22:30	2:51:24	NA	NA
	P	0.924	0.603	0.570	0.950	0.921	0.992	NA	NA
20	D	28.3	41.6	46.3	48.5	77.0	26.6	NA	NA
Period 1	T	12:47:10	13:00:02	14:49:05	12:26:44	13:08:27	6:43:52	NA	NA
	P	0.731	0.832	0.874	0.625	0.730	0.671	NA	NA
20	D	31.4	31.9	31.5	142.9	246.0	110.7	NA	NA
Period 2	T	8:54:52	13:07:53	12:57:18	13:16:10	11:42:49	12:04:10	NA	NA
	P	0.658	0.872	0.712	0.762	0.810	0.734	NA	NA
21	D	21.3	28.7	31.7	23.1	48.9	13.5	29.1	NA
Period 1	T	12:43:30	13:38:05	9:09:33	9:39:17	13:43:14	10:43:24	13:08:41	NA
	P	0.465	0.644	0.791	0.633	0.414	0.447	0.659	NA
21	D	21.6	18.0	2.1	14.7	29.0	5.5	15.0	NA
Period 2	T	13:08:23	13:42:15	10:54:19	10:18:56	11:50:53	11:06:44	8:37:24	NA
	P	0.514	0.383	0.089	0.538	0.392	0.511	0.465	NA
22	D	37.9	28.9	49.4	62.3	28.1	18.1	NA	NA
Period 1	T	12:56:31	14:16:20	13:22:23	13:43:03	12:04:19	3:49:18	NA	NA
	P	0.777	0.826	0.796	0.733	0.766	0.835	NA	NA
22	D	22.6	40.0	28.0	33.6	67.0	20.8	13.0	NA
Period 2	T	10:32:49	14:05:53	11:38:53	13:14:48	11:31:07	12:37:37	9:42:48	NA
	P	0.626	0.605	0.722	0.635	0.816	0.669	0.536	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
23	D	24.4	33.6	30.7	14.3	20.1	27.2	1.1	NA
Period 1	T	10:39:23	13:40:51	12:38:45	12:03:21	13:13:47	9:27:57	0:51:30	NA
	P	0.649	0.638	0.694	0.955	0.830	0.772	0.879	NA
23	D	11.5	31.3	28.2	33.6	17.7	39.2	6.9	NA
Period 2	T	7:07:49	12:50:21	14:02:56	14:15:56	12:01:51	14:08:39	7:37:10	NA
	P	0.717	0.476	0.815	0.537	0.875	0.609	0.723	NA
24	D	56.7	78.6	70.9	61.9	44.9	49.0	1.2	NA
Period 1	T	12:03:24	15:06:46	16:17:31	13:54:43	12:50:04	13:11:45	7:51:55	NA
	P	0.595	0.775	0.802	0.827	0.457	0.387	0.059	NA
24	D	52.8	39.2	41.1	36.8	24.1	39.8	24.0	NA
Period 2	T	10:32:12	6:44:44	9:14:05	10:06:04	6:45:07	17:44:49	5:09:58	NA
	P	0.690	0.764	0.936	0.676	0.730	0.317	0.369	NA
25	D	60.5	34.8	51.4	50.5	33.7	NA	NA	NA
Period 1	T	22:38:25	11:16:15	10:42:42	12:26:58	8:30:54	NA	NA	NA
	P	0.925	0.925	0.806	0.790	0.600	NA	NA	NA
25	D	30.9	332.6	343.7	46.3	29.3	3.5	NA	NA
Period 2	T	10:02:53	14:06:06	12:29:59	11:35:52	9:23:57	1:07:05	NA	NA
	P	0.902	0.956	0.967	0.905	0.931	0.994	NA	NA
26	D	47.9	163.0	355.4	21.1	23.5	27.3	0.9	NA
Period 1	T	9:49:31	11:29:41	12:25:19	10:38:31	11:56:58	7:29:24	1:13:23	NA
	P	0.805	0.914	0.958	0.708	0.883	0.744	0.844	NA
26	D	30.8	37.2	25.2	7.3	16.9	47.8	43.4	26.5
Period 2	T	7:28:45	11:28:21	9:30:41	8:53:35	10:22:17	7:34:45	12:33:48	9:50:14
	P	0.731	0.488	0.760	0.884	0.872	0.702	0.745	0.521
27	D	24.6	70.3	38.0	20.4	6.0	13.2	NA	NA
Period 1	T	5:28:33	18:12:24	12:44:16	5:56:39	2:23:36	5:18:14	NA	NA
	P	0.627	0.766	0.766	0.677	0.753	0.603	NA	NA
27	D	7.2	10.5	1.6	3.4	4.0	13.6	NA	NA
Period 2	T	4:30:12	5:04:47	2:21:04	2:15:29	1:20:23	1:34:00	NA	NA
	P	0.347	0.442	0.104	0.421	0.734	0.232	NA	NA
28	D	12.8	23.3	26.5	34.1	53.5	40.9	23.0	NA
Period 1	T	8:10:56	5:58:37	10:12:20	9:54:40	7:00:15	11:18:17	6:22:54	NA
	P	0.404	0.778	0.728	0.514	0.591	0.403	0.783	NA
28	D	6.4	64.5	13.1	23.8	11.1	20.4	NA	NA
Period 2	T	1:18:00	16:06:50	4:27:01	8:37:26	5:34:05	8:10:49	NA	NA
	P	0.874	0.471	0.567	0.542	0.309	0.514	NA	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
29	D	26.6	9.1	65.7	65.3	65.6	25.9	13.0	14.6
Period 1	T	10:47:58	12:40:29	13:10:29	12:22:30	2:23:20	9:19:35	6:38:41	3:55:58
	P	0.677	0.344	0.752	0.455	0.974	0.469	0.330	0.889
29	D	0.1	18.9	10.7	26.4	10.2	17.1	23.7	72.8
Period 2	T	0:18:28	12:35:41	11:34:33	5:39:04	7:47:39	8:32:31	9:51:30	8:15:36
	P	0.638	0.766	0.199	0.923	0.656	0.633	0.613	0.652
30	D	0.2	4.8	7.8	43.8	37.6	67.8	94.2	54.9
Period 1	T	0:15:43	6:48:17	9:51:30	10:46:58	12:43:53	11:50:47	10:59:45	7:44:59
	P	0.972	0.306	0.316	0.685	0.580	0.695	0.432	0.612
30	D	0.1	39.2	7.9	1.8	28.5	52.1	50.1	5.0
Period 2	T	2:51:38	14:17:24	8:21:00	5:00:18	14:06:41	11:12:17	11:34:27	1:32:25
	P	0.166	0.223	0.351	0.183	0.465	0.468	0.280	0.484
31	D	11.0	9.6	9.9	27.2	58.4	28.1	NA	NA
Period 1	T	11:43:03	14:19:32	13:22:02	13:27:18	13:15:24	10:27:29	NA	NA
	P	0.699	0.706	0.465	0.783	0.797	0.897	NA	NA
31	D	15.7	58.4	17.9	52.7	24.2	17.7	17.7	NA
Period 2	T	7:26:18	11:04:08	13:55:58	14:10:17	13:00:10	12:11:31	12:11:31	NA
	P	0.756	0.758	0.669	0.807	0.748	0.723	0.723	NA
32	D	29.5	50.2	59.8	38.2	56.1	40.4	51.4	34.4
Period 1	T	6:57:57	15:02:18	9:45:11	13:30:07	10:19:54	8:31:47	10:53:35	6:06:51
	P	0.771	0.582	0.802	0.685	0.704	0.430	0.411	0.960
32	D	6.8	51.9	6.6	66.7	45.8	33.5	84.0	165.2
Period 2	T	1:46:02	13:46:13	6:03:06	10:47:38	12:31:44	12:51:58	11:56:09	9:00:40
	P	0.840	0.540	0.397	0.729	0.730	0.821	0.731	0.814
33	D	18.1	25.2	59.4	137.7	49.7	46.9	107.7	3.2
Period 1	T	9:12:52	10:40:14	11:09:57	16:38:39	11:35:23	8:06:21	12:10:42	1:02:13
	P	0.683	0.317	0.851	0.622	0.572	0.749	0.567	0.927
33	D	28.7	29.0	35.2	137.1	75.6	3.9	85.8	1.3
Period 2	T	6:44:08	10:37:29	10:49:32	10:33:00	8:40:46	1:28:28	16:06:22	0:12:12
	P	0.918	0.472	0.631	0.776	0.785	0.787	0.556	0.964
34	D	76.5	49.0	55.8	27.2	23.9	9.0	NA	NA
Period 1	T	9:06:51	15:56:40	11:05:04	13:44:09	13:59:42	4:00:26	NA	NA
	P	0.827	0.730	0.711	0.919	0.844	0.882	NA	NA
34	D	4.5	34.4	25.7	35.9	362.1	8.3	59.0	NA
Period 2	T	2:17:32	16:12:29	12:46:47	9:33:54	14:20:23	3:51:11	10:59:31	NA
	P	0.813	0.789	0.906	0.538	0.815	0.857	0.786	NA
35	D	20.0	25.0	23.2	0.1	NA	NA	NA	NA
Period 1	T	14:59:23	24:00:00	24:00:00	5:49:37	NA	NA	NA	NA
	P	0.709	0.673	0.647	1.0	NA	NA	NA	NA
35	D	24.4	100.3	445.2	63.4	11.1	37.0	3.3	NA
Period 2	T	10:10:44	10:50:34	13:01:39	14:04:23	10:12:54	10:18:35	2:31:38	NA
	P	0.454	0.804	0.838	0.643	0.646	0.595	0.814	NA

Appendix B

Number of clusters and the proportion of time spent in the clusters

Table B.1: Number of Clusters and Proportion in Cluster

Subject	Statistic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
1	Clusters	4	4	5	3	6	4	3	NA
Period 1	Proportion	0.765	0.823	0.617	0.628	0.427	0.665	0.837	NA
1	Clusters	6	9	3	4	6	1	NA	NA
Period 2	Proportion	0.591	0.383	0.801	0.624	0.494	0.830	NA	NA
2	Clusters	5	2	4	1	NA	NA	NA	NA
Period 1	Proportion	0.820	0.959	0.923	0.986	NA	NA	NA	NA
2	Clusters	4	5	5	5	1	NA	NA	NA
Period 2	Proportion	0.650	0.671	0.651	0.879	0.509	NA	NA	NA
3	Clusters	1	6	1	3	5	2	1	2
Period 1	Proportion	0.999	0.599	0.988	0.871	0.856	0.973	0.981	0.662
3	Clusters	1	4	1	2	4	2	5	1
Period 2	Proportion	1.0	0.614	0.987	0.497	0.626	0.934	0.591	0.425
4	Clusters	1	7	4	5	3	8	NA	NA
Period 1	Proportion	0.925	0.786	0.507	0.716	0.892	0.839	NA	NA
4	Clusters	2	3	5	7	5	3	2	NA
Period 2	Proportion	0.824	0.845	0.745	0.746	0.870	0.812	0.689	NA

Subject	Statistic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
5	Clusters	2	6	2	3	3	1	NA	NA
Period 1	Proportion	0.850	0.768	0.906	0.611	0.719	0.991	NA	NA
5	Clusters	3	9	5	5	5	2	NA	NA
Period 2	Proportion	0.841	0.784	0.918	0.658	0.822	0.919	NA	NA
6	Clusters	8	6	7	8	5	NA	NA	NA
Period 1	Proportion	0.755	0.813	0.826	0.681	0.836	NA	NA	NA
6	Clusters	6	6	3	NA	NA	NA	NA	NA
Period 2	Proportion	0.755	0.802	0.652	NA	NA	NA	NA	NA
7	Clusters	3	2	4	6	3	4	4	NA
Period 1	Proportion	0.477	0.382	0.730	0.254	0.824	0.569	0.363	NA
7	Clusters	3	3	3	5	5	3	3	2
Period 2	Proportion	0.639	0.374	0.842	0.308	0.610	0.736	0.204	0.698
8	Clusters	1	4	6	1	1	3	NA	NA
Period 1	Proportion	0.916	0.875	0.683	0.879	0.817	0.416	NA	NA
8	Clusters	1	4	4	1	6	2	NA	NA
Period 2	Proportion	0.857	0.919	0.882	0.849	0.468	0.917	NA	NA
9	Clusters	2	3	3	1	1	1	NA	NA
Period 1	Proportion	0.863	0.887	0.900	1.0	1.0	1.0	NA	NA
9	Clusters	3	1	3	2	1	2	4	1
Period 2	Proportion	0.853	0.999	0.875	0.831	0.995	0.725	0.815	0.649
10	Clusters	3	2	1	3	5	1	1	3
Period 1	Proportion	0.908	0.944	0.993	0.753	0.876	0.981	0.946	0.599
11	Clusters	3	6	6	6	4	4	3	1
Period 1	Proportion	0.765	0.763	0.794	0.780	0.685	0.357	0.802	0.889
11	Clusters	8	4	6	7	4	4	2	NA
Period 2	Proportion	0.621	0.879	0.783	0.755	0.453	0.713	0.357	NA
12	Clusters	3	3	3	4	4	3	1	0
Period 1	Proportion	0.550	0.537	0.489	0.357	0.464	0.683	0.372	0
12	Clusters	4	4	3	3	2	3	3	2
Period 2	Proportion	0.522	0.467	0.656	0.085	0.567	0.475	0.236	0.321
13	Clusters	3	3	2	1	NA	NA	NA	NA
Period 1	Proportion	0.938	0.947	0.952	1.0	NA	NA	NA	NA
13	Clusters	1	4	5	6	1	4	NA	NA
Period 2	Proportion	0.892	0.764	0.783	0.999	0.996	0.796	NA	NA
14	Clusters	3	5	2	3	7	1	NA	NA
Period 1	Proportion	0.861	0.793	0.576	0.803	0.113	0.964	NA	NA
14	Clusters	4	3	3	1	4	1	4	NA
Period 2	Proportion	0.548	0.909	0.881	0.996	0.623	0.990	0.903	NA
15	Clusters	6	3	4	4	5	2	NA	NA
Period 1	Proportion	0.451	0.620	0.450	0.340	0.505	0.109	NA	NA

Subject	Statistic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
16	Clusters	6	2	3	6	9	4	NA	NA
Period 1	Proportion	0.824	0.598	0.651	0.829	0.341	0.746	NA	NA
16	Clusters	3	3	6	4	9	5	2	NA
Period 2	Proportion	0.472	0.807	0.650	0.888	0.368	0.824	0.661	NA
17	Clusters	4	4	3	2	1	NA	NA	NA
Period 1	Proportion	0.753	0.726	0.848	0.897	0.922	NA	NA	NA
17	Clusters	1	3	3	2	3	NA	NA	NA
Period 2	Proportion	0.807	0.525	0.410	0.557	0.432	NA	NA	NA
18	Clusters	2	1	2	5	5	NA	NA	NA
Period 1	Proportion	0.947	0.943	0.945	0.953	0.537	NA	NA	NA
18	Clusters	4	4	3	1	3	2	NA	NA
Period 2	Proportion	0.930	0.987	0.979	1.00	0.963	0.954	NA	NA
19	Clusters	2	4	3	6	3	NA	NA	NA
Period 1	Proportion	0.899	0.822	0.927	0.782	0.937	NA	NA	NA
19	Clusters	6	7	8	6	6	2	NA	NA
Period 2	Proportion	0.895	0.615	0.681	0.834	0.785	0.282	NA	NA
20	Clusters	4	10	5	6	4	4	NA	NA
Period 1	Proportion	0.821	0.630	0.767	0.682	0.651	0.578	NA	NA
20	Clusters	5	7	4	4	6	9	NA	NA
Period 2	Proportion	0.810	0.731	0.748	0.673	0.508	0.305	NA	NA
21	Clusters	4	4	4	3	7	2	4	NA
Period 1	Proportion	0.603	0.794	0.794	0.657	0.659	0.871	0.831	NA
21	Clusters	6	5	1	2	4	1	1	NA
Period 2	Proportion	0.727	0.722	0.205	0.718	0.535	0.993	0.945	NA
22	Clusters	4	5	6	5	7	3	NA	NA
Period 1	Proportion	0.897	0.891	0.753	0.767	0.819	0.708	NA	NA
22	Clusters	2	2	4	3	4	3	2	NA
Period 2	Proportion	0.891	0.856	0.841	0.842	0.715	0.838	0.857	NA
23	Clusters	1	2	4	1	3	4	1	NA
Period 1	Proportion	0.831	0.694	0.846	0.962	0.907	0.693	0.959	NA
23	Clusters	2	2	4	4	1	11	1	NA
Period 2	Proportion	0.869	0.564	0.871	0.553	0.888	0.605	0.903	NA
24	Clusters	3	4	1	1	4	4	1	NA
Period 1	Proportion	0.394	0.526	0.313	0.543	0.447	0.195	0.330	NA
24	Clusters	2	3	1	2	2	3	3	NA
Period 2	Proportion	0.443	0.345	0.514	0.519	0.438	0.754	0.346	NA
25	Clusters	3	6	3	7	6	NA	NA	NA
Period 1	Proportion	0.756	0.699	0.633	0.674	0.454	NA	NA	NA
25	Clusters	5	4	6	6	3	1	NA	NA
Period 2	Proportion	0.733	0.691	0.540	0.656	1.0	0.952	NA	NA

Subject	Statistic	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
26	Clusters	7	5	5	2	3	6	1	NA
Period 1	Proportion	0.540	0.624	0.614	0.772	0.848	0.630	0.983	NA
26	Clusters	8	5	7	3	4	4	8	7
Period 2	Proportion	0.688	0.446	0.695	0.887	0.856	0.546	0.651	0.572
27	Clusters	3	5	5	2	3	2	NA	NA
Period 1	Proportion	0.585	0.581	0.628	0.622	0.433	0.577	NA	NA
27	Clusters	3	4	1	2	2	1	NA	NA
Period 2	Proportion	0.309	0.284	0.509	0.580	0.391	0.081	NA	NA
28	Clusters	2	2	3	5	2	6	2	NA
Period 1	Proportion	0.345	0.464	0.665	0.579	0.409	0.355	0.555	NA
28	Clusters	1	5	1	2	3	3	NA	NA
Period 2	Proportion	0.605	0.289	0.528	0.470	0.323	0.302	NA	NA
29	Clusters	3	2	4	4	3	5	3	4
Period 1	Proportion	0.721	0.425	0.818	0.860	0.450	0.304	0.318	0.533
29	Clusters	1	4	4	5	2	3	4	9
Period 2	Proportion	0.888	0.844	0.259	0.545	0.827	0.796	0.819	0.586
30	Clusters	1	2	2	6	4	6	3	3
Period 1	Proportion	1.0	0.916	0.942	0.722	0.726	0.629	0.650	0.657
30	Clusters	1	4	1	1	4	4	5	5
Period 2	Proportion	1.0	0.350	0.938	0.940	0.864	0.724	0.602	0.775
31	Clusters	1	1	3	3	5	1	NA	NA
Period 1	Proportion	0.932	0.971	0.888	0.893	0.825	0.869	NA	NA
31	Clusters	5	2	1	3	1	4	4	NA
Period 2	Proportion	0.858	0.886	0.920	0.825	0.849	0.870	0.870	NA
32	Clusters	2	9	2	3	3	5	5	3
Period 1	Proportion	0.688	0.725	0.562	0.857	0.523	0.481	0.824	0.793
32	Clusters	1	3	1	3	2	4	6	5
Period 2	Proportion	0.546	0.570	0.849	0.706	0.626	0.824	0.588	0.526
33	Clusters	2	2	3	4	5	4	6	1
Period 1	Proportion	0.655	0.307	0.665	0.654	0.374	0.488	0.675	0.675
33	Clusters	2	4	5	5	4	2	4	0
Period 2	Proportion	0.598	0.623	0.695	0.415	0.416	0.390	0.597	0
34	Clusters	3	9	5	2	3	2	NA	NA
Period 1	Proportion	0.786	0.804	0.762	0.912	0.913	0.930	NA	NA
34	Clusters	1	3	2	4	10	5	6	NA
Period 2	Proportion	0.675	0.908	0.868	0.790	0.609	0.708	0.613	NA
35	Clusters	3	1	3	1	NA	NA	NA	NA
Period 1	Proportion	0.932	0.864	0.913	1.0	NA	NA	NA	NA
35	Clusters	2	4	3	4	1	2	1	NA
Period 2	Proportion	0.511	0.821	0.571	0.716	0.866	0.832	0.924	NA

Appendix C

Total distance travelled: unfiltered (km)(D), Distance of smoothed series (D*), Lower bound on distance (D**), Length of time of recorded series (hours) (T)

Table C.1: Distance Measurements

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
1	D	73.5	10.2	30.7	29.0	20.3	27.7	25.8	NA
	D*	70.8	9.8	27.4	25.4	19.5	25.1	20.9	NA
	D**	63.2	4.5	22.0	18.2	16.3	14.0	9.5	NA
Period 1	T	10:14:32	4:53:57	10:56:30	8:36:19	8:6:31	10:39:15	8:21:11	NA
1	D	38.1	30.9	23.4	17.8	55.5	5.0	NA	NA
	D*	34.3	27.6	21.9	16.7	50.3	4.6	NA	NA
	D**	24.4	24.9	5.8	8.3	40.9	2.4	NA	NA
Period 2	T	13:47:9	7:58:41	12:13:31	11:58:43	12:14:46	1:26:48	NA	NA
2	D	43.4	21.6	53.8	2.5	NA	NA	NA	NA
	D*	43.1	19.0	51.6	2.4	NA	NA	NA	NA
	D**	39.7	12.7	42.0	0.2	NA	NA	NA	NA
Period 1	T	13:56:05	24:00:00	22:07:59	4:28:15	NA	NA	NA	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
2	D	207.8	93.4	172.3	31.9	4.0	NA	NA	NA
	D*	207.0	92.7	172.3	31.7	4.1	NA	NA	NA
	D**	201.7	88.4	170.9	23.6	3.8	NA	NA	NA
Period 2	T	11:21:40	9:31:54	11:56:12	13:42:03	0:38:52	NA	NA	NA
3	D	0.7	270.5	5.3	17.4	33.6	6.4	4.5	8.7
	D*	0.7	269.8	5.2	17.5	32.4	6.3	4.5	8.5
	D**	0.2	268.1	0.7	15.0	28.7	5.2	0.2	7.5
Period 1	T	1:00:41	10:59:33	8:53:45	9:20:32	10:14:06	14:32:00	10:30:15	7:28:48
3	D	0.8	256.6	7.4	15.4	99.5	18.4	256.5	12.5
	D*	0.8	256.9	7.3	15.1	99.1	18.4	256.4	12.1
	D**	0.1	256.2	0.0	12.9	95.2	17.5	254.7	11.2
Period 2	T	1:29:15	8:22:27	8:13:20	7:46:25	13:31:29	13:47:42	8:40:43	6:28:48
4	D	7.3	68.4	109.3	91.3	33.4	43.7	NA	NA
	D*	7.2	62.7	100.6	84.8	32.4	43.2	NA	NA
	D**	3.4	55.0	90.3	74.6	24.9	28.3	NA	NA
Period 1	T	6:27:39	13:51:03	15:48:51	14:10:26	14:07:27	16:02:30	NA	NA
4	D	11.9	31.5	79.9	58.7	36.0	28.5	33.4	NA
	D*	11.8	30.2	75.0	55.9	33.2	23.7	33.3	NA
	D**	8.6	21.7	64.4	47.8	26.9	17.2	31.3	NA
Period 2	T	6:13:42	15:48:44	15:30:46	15:49:59	13:07:12	10:12:45	9:04:17	NA
5	D	31.4	148.2	39.9	53.3	57.9	4.1	NA	NA
	D*	31.2	147.7	39.3	44.7	49.5	3.7	NA	NA
	D**	28.5	139.6	34.4	30.9	43.8	1.5	NA	NA
Period 1	T	9:21:15	15:01:28	15:45:00	16:05:30	11:47:26	15:55:40	NA	NA
5	D	29.6	77.1	63.1	75.2	28.5	17.4	NA	NA
	D*	29.2	76.3	63.0	71.7	25.8	17.3	NA	NA
	D**	24.0	70.5	59.3	65.2	21.2	12.1	NA	NA
Period 2	T	8:25:48	16:56:27	24:11:59	12:58:38	14:11:40	7:41:34	NA	NA
6	D	46.4	103.4	74.9	37.8	47.2	NA	NA	NA
	D*	46.0	101.2	74.3	37.1	46.3	NA	NA	NA
	D**	42.2	97.3	65.5	34.6	42.4	NA	NA	NA
Period 1	T	13:11:31	23:07:32	14:36:08	12:36:16	10:42:32	NA	NA	NA
6	D	73.6	70.0	11.5	NA	NA	NA	NA	NA
	D*	73.0	69.7	9.2	NA	NA	NA	NA	NA
	D**	69.6	67.8	8.0	NA	NA	NA	NA	NA
Period 2	T	11:54:31	13:12:36	2:22:20	NA	NA	NA	NA	NA
7	D	20.4	26.7	57.4	17.8	17.1	20.6	23.8	NA
	D*	17.8	21.4	54.6	15.4	16.4	18.8	19.2	NA
	D**	10.3	16.6	49.4	14.3	7.5	15.2	17.6	NA
Period 1	T	9:48:47	7:34:19	13:45:41	5:46:23	7:10:42	10:40:38	7:59:10	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
7	D	16.7	28.2	3.9	31.4	20.8	12.0	14.4	14.9
	D*	15.9	23.4	3.9	28.1	19.2	11.2	13.0	14.4
	D**	11.8	19.9	3.0	26.3	11.0	10.4	11.7	7.4
Period 2	T	5:59:10	8:38:13	4:27:07	9:06:45	7:46:16	8:01:30	6:15:45	5:26:05
8	D	13.4	24.9	37.3	24.4	38.1	28.3	NA	NA
	D*	12.6	23.6	34.9	21.3	34.3	26.4	NA	NA
	D**	3.3	12.7	20.7	9.0	15.3	21.7	NA	NA
Period 1	T	9:33:29	15:12:18	9:18:01	11:38:12	12:01:02	8:29:04	NA	NA
8	D	23.4	26.9	31.0	18.5	32.9	8.7	NA	NA
	D*	22.0	25.9	29.5	17.7	31.7	8.5	NA	NA
	D**	10.6	9.7	13.3	6.2	21.2	1.4	NA	NA
Period 2	T	8:29:04	7:00:55	15:00:57	14:33:47	11:32:14	5:30:41	NA	NA
9	D	29.6	17.6	127.9	0.3	0.7	0.2	NA	NA
	D*	28.3	16.4	128.0	0.3	0.7	0.2	NA	NA
	D**	20.8	8.7	125.5	0.0	0.0	0.0	NA	NA
Period 1	T	12:13:56	23:27:07	16:15:25	15:19:07	24:00:00	5:48:57	NA	NA
9	D	26.5	0.8	7.4	12.2	6.7	11.9	14.2	8.3
	D*	24.8	0.7	6.7	12.0	6.6	11.9	14.0	8.0
	D**	10.0	0.0	5.6	10.5	0.1	10.6	13.0	7.9
Period 2	T	11:00:57	19:42:53	21:08:31	13:10:33	12:21:53	12:16:47	4:35:31	1:33:37
10	D	10.5	28.0	3.2	37.6	37.5	8.1	19.9	50.5
	D*	9.4	27.8	3.1	37.4	36.8	7.4	18.3	50.1
	D**	6.3	21.9	0.1	30.5	32.7	0.5	5.1	46.1
Period 1	T	8:45:08	15:10:41	3:58:31	11:58:31	11:07:14	13:20:16	10:49:58	8:07:38
11	D	58.6	53.6	73.2	37.0	74.5	40.3	13.0	1.5
	D*	56.8	53.0	72.7	36.8	72.0	38.6	12.2	1.4
	D**	49.3	50.2	67.8	30.3	63.4	37.4	7.4	0.8
Period 1	T	13:32:35	23:40:07	13:42:32	12:23:16	12:16:16	15:20:27	16:49:20	0:58:46
11	D	131.9	109.2	125.1	76.9	82.9	43.0	19.4	NA
	D*	131.7	108.6	125.0	75.6	78.9	42.1	18.3	NA
	D**	129.7	104.3	120.8	68.1	69.9	33.8	18.2	NA
Period 2	T	13:16:57	15:34:33	9:48:45	16:39:03	13:20:47	13:53:59	5:08:56	NA
12	D	44.0	79.8	40.5	47.8	43.4	48.5	26.0	1.4
	D*	36.0	63.6	37.2	38.0	39.9	40.9	17.6	0.4
	D**	29.5	51.2	32.2	32.2	36.2	30.7	12.2	0.2
Period 1	T	13:09:08	14:13:12	11:54:23	15:46:24	9:21:46	10:51:13	13:17:53	0:10:23
12	D	49.9	71.5	89.7	36.0	19.5	35.5	53.5	59.6
	D*	43.6	67.2	73.7	32.5	17.5	30.8	46.9	54.3
	D**	40.7	60.9	56.9	31.6	16.0	24.8	42.6	46.8
Period 2	T	13:19:56	12:03:12	15:05:11	12:43:46	15:18:15	10:32:59	14:45:38	27:19:56

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
13	D	11.7	15.2	28.0	0.2	NA	NA	NA	NA
	D*	11.2	14.3	27.7	0.2	NA	NA	NA	NA
	D**	6.2	6.5	20.9	0.0	NA	NA	NA	NA
Period 1	T	13:46:37	24:00:00	24:00:00	3:56:10	NA	NA	NA	NA
13	D	15.2	34.8	32.2	65.7	3.1	15.7	NA	NA
	D*	13.5	33.8	31.0	63.7	3.0	15.0	NA	NA
	D**	5.9	24.9	24.0	58.1	0.4	8.1	NA	NA
Period 2	T	11:54:24	14:33:38	11:35:19	13:42:33	11:28:17	6:20:00	NA	NA
14	D	28.7	54.5	22.3	26.5	597.9	0.6	NA	NA
	D*	28.3	53.3	20.1	25.4	596.3	0.6	NA	NA
	D**	23.2	42.2	14.5	16.6	595.4	0.3	NA	NA
Period 1	T	10:34:37	12:22:35	11:48:49	8:22:47	13:31:57	0:54:40	NA	NA
14	D	15.1	20.2	14.8	3.1	147.3	5.7	20.3	NA
	D*	15.0	20.0	14.5	3.0	146.1	5.4	19.7	NA
	D**	13.5	17.0	9.8	0.6	144.6	0.9	14.9	NA
Period 2	T	6:29:37	9:56:09	9:23:55	9:47:10	8:14:14	6:19:21	6:50:59	NA
15	D	41.4	24.3	19.6	37.3	26.8	8.0	NA	NA
	D*	34.1	21.7	16.7	28.0	23.2	7.1	NA	NA
	D**	32.2	18.3	15.6	26.6	17.1	6.5	NA	NA
Period 1	T	10:04:33	9:40:31	6:08:32	9:44:00	6:43:04	6:59:39	NA	NA
16	D	42.4	23.3	53.3	33.4	329.3	107.3	NA	NA
	D*	41.4	23.3	51.2	32.4	329.6	105.0	NA	NA
	D**	32.8	21.1	42.1	23.5	327.0	95.6	NA	NA
Period 1	T	11:40:36	4:21:13	8:00:04	13:56:24	13:05:01	11:41:22	NA	NA
16	D	35.0	24.0	47.5	26.0	336.8	105.0	23.8	NA
	D*	32.5	23.4	46.5	25.6	334.7	104.7	23.5	NA
	D**	28.8	13.5	42.8	16.1	331.9	100.7	20.5	NA
Period 2	T	7:54:47	10:57:35	10:13:48	10:09:12	12:17:35	13:41:12	3:21:52	NA
17	D	20.2	22.4	13.3	15.3	1.3	NA	NA	NA
	D	19.0	20.7	12.5	14.8	1.1	NA	NA	NA
	D**	12.3	14.4	9.0	10.2	0.6	NA	NA	NA
Period 1	T	9:26:28	10:08:20	12:10:29	19:55:27	0:28:36	NA	NA	NA
17	D	0.5	12.7	18.6	9.0	27.4	NA	NA	NA
	D*	0.5	11.8	16.5	8.1	23.6	NA	NA	NA
	D**	0.4	10.0	11.8	7.2	20.4	NA	NA	NA
Period 2	T	4:15:49	4:47:09	7:32:38	3:18:05	7:56:14	NA	NA	NA
18	D	10.5	11.6	24.5	30.8	50.9	NA	NA	NA
	D*	10.4	11.2	23.8	29.6	49.1	NA	NA	NA
	D**	4.5	3.8	10.5	19.2	39.0	NA	NA	NA
Period 1	T	11:50:19	13:21:45	14:31:05	13:02:20	13:19:08	NA	NA	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
18	D	26.6	26.1	23.3	14.3	21.9	7.9	NA	NA
	D*	24.0	25.6	22.8	14.0	19.6	7.7	NA	NA
	D**	18.5	14.6	10.4	0.3	14.3	4.6	NA	NA
Period 2	T	10:09:55	15:09:56	12:56:15	14:24:00	12:23:04	2:23:19	NA	NA
19	D	14.2	21.3	14.2	35.5	14.2	NA	NA	NA
	D*	13.5	20.4	14.1	33.2	13.4	NA	NA	NA
	D**	8.8	11.8	6.0	28.2	7.6	NA	NA	NA
Period 1	T	12:41:47	11:57:18	12:09:49	13:52:52	18:39:28	NA	NA	NA
19	D	24.5	31.7	32.7	28.1	63.5	33.0	NA	NA
	D*	23.9	30.9	32.4	26.9	62.9	32.2	NA	NA
	D**	16.1	24.2	29.0	19.0	57.2	31.4	NA	NA
Period 2	T	12:57:11	11:37:28	12:35:55	13:26:52	14:22:30	2:51:24	NA	NA
20	D	28.3	41.6	46.3	48.5	77.0	26.6	NA	NA
	D*	27.0	41.4	45.6	47.8	75.6	25.5	NA	NA
	D**	20.1	38.6	31.6	42.3	70.7	24.9	NA	NA
Period 1	T	12:47:10	13:00:02	14:49:05	12:26:44	13:08:27	6:43:52	NA	NA
20	D	31.4	31.9	31.5	142.9	246.0	110.7	NA	NA
	D*	30.6	30.0	30.9	141.5	245.3	108.8	NA	NA
	D**	25.7	26.3	26.5	131.0	241.5	104.8	NA	NA
Period 2	T	8:54:52	13:07:53	12:57:18	13:16:10	11:42:49	12:04:10	NA	NA
21	D	21.3	28.7	31.7	23.1	48.9	13.5	29.1	NA
	D*	20.8	27.5	31.6	21.0	46.5	13.0	27.6	NA
	D**	16.2	19.0	29.5	13.2	42.5	3.0	15.6	NA
Period 1	T	12:43:30	13:38:05	9:09:33	9:39:17	13:43:14	10:43:24	13:08:41	NA
21	D	21.6	18.0	2.1	14.7	29.0	5.5	15.0	NA
	D*	21.2	17.4	2.1	13.6	28.3	5.4	11.1	NA
	D**	17.6	13.9	1.1	7.7	23.1	0.3	6.1	NA
Period 2	T	13:08:23	13:42:15	10:54:19	10:18:56	11:50:53	11:06:44	8:37:24	NA
22	D	37.9	28.9	49.4	62.3	28.1	18.1	NA	NA
	D*	34.7	27.4	46.8	60.8	27.0	17.9	NA	NA
	D**	24.9	17.4	38.2	53.6	20.0	14.6	NA	NA
Period 1	T	12:56:31	14:16:20	13:22:23	13:43:03	12:04:19	3:49:18	NA	NA
22	D	22.6	40.0	28.0	33.6	67.0	20.8	13.0	NA
	D*	20.6	38.9	27.2	32.8	66.1	20.1	12.7	NA
	D**	16.2	32.2	20.0	26.5	61.0	11.9	8.5	NA
Period 2	T	10:32:49	14:05:53	11:38:53	13:14:48	11:31:07	12:37:37	9:42:48	NA
23	D	24.4	33.6	30.7	14.3	20.1	27.2	1.1	NA
	D*	20.3	31.8	28.6	13.3	18.7	25.9	1.0	NA
	D**	7.1	19.5	17.1	1.9	5.7	19.2	0.2	NA
Period 1	T	10:39:23	13:40:51	12:38:45	12:03:21	13:13:47	9:27:57	0:51:30	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
23	D	11.5	31.3	28.2	33.6	17.7	39.2	6.9	NA
	D*	11.2	30.6	26.3	33.1	15.9	38.1	6.0	NA
	D**	2.6	23.3	15.4	26.0	5.0	32.2	1.1	NA
Period 2	T	7:07:49	12:50:21	14:02:56	14:15:56	12:01:51	14:08:39	7:37:10	NA
24	D	56.7	78.6	70.9	61.9	44.9	49.0	1.2	NA
	D*	43.9	60.5	56.5	46.3	37.5	44.6	1.0	NA
	D**	38.8	53.6	40.1	34.6	34.4	40.8	0.8	NA
Period 1	T	12:03:24	15:06:46	16:17:31	13:54:43	12:50:04	13:11:45	7:51:55	NA
24	D	52.8	39.2	41.1	36.8	24.1	39.8	24.0	NA
	D*	41.1	24.0	30.2	29.3	20.9	35.6	21.9	NA
	D**	33.3	19.8	21.3	23.0	17.5	29.0	20.0	NA
Period 2	T	10:32:12	6:44:44	9:14:05	10:06:04	6:45:07	17:44:49	5:09:58	NA
25	D	60.5	34.8	51.4	50.5	33.7	NA	NA	NA
	D	51.4	30.9	49.5	48.2	29.3	NA	NA	NA
	D**	24.8	19.9	36.5	38.1	23.5	NA	NA	NA
Period 1	T	22:38:25	11:16:15	10:42:42	12:26:58	8:30:54	NA	NA	NA
25	D	30.9	332.6	343.7	46.3	29.3	3.5	NA	NA
	D*	28.2	331.7	340.5	43.9	25.0	3.0	NA	NA
	D**	15.2	324.5	330.1	31.7	15.8	0.7	NA	NA
Period 2	T	10:02:53	14:06:06	12:29:59	11:35:52	9:23:57	1:07:05	NA	NA
26	D	47.9	163.0	355.4	21.1	23.5	27.3	0.9	NA
	D*	40.1	161.7	355.2	20.2	22.7	26.7	0.8	NA
	D**	37.0	159.3	351.7	17.6	19.6	23.9	0.2	NA
Period 1	T	9:49:31	11:29:41	12:25:19	10:38:31	11:56:58	7:29:24	1:13:23	NA
26	D	30.8	37.2	25.2	7.3	16.9	47.8	43.4	26.5
	D*	30.6	35.7	24.2	6.7	16.5	44.5	41.8	25.8
	D**	29.2	33.8	20.2	5.0	12.8	42.0	37.3	23.3
Period 2	T	7:28:45	11:28:21	9:30:41	8:53:35	10:22:17	7:34:45	12:33:48	9:50:14
27	D	24.6	70.3	38.0	20.4	6.0	13.2	NA	NA
	D*	20.6	61.6	31.6	15.9	5.4	10.8	NA	NA
	D**	16.4	39.9	16.6	11.5	4.0	6.6	NA	NA
Period 1	T	5:28:33	18:12:24	12:44:16	5:56:39	2:23:36	5:18:14	NA	NA
27	D	7.2	10.5	1.6	3.4	4.0	13.6	NA	NA
	D*	6.0	8.5	1.4	3.0	3.6	13.0	NA	NA
	D**	5.6	7.5	1.4	2.7	3.5	13.0	NA	NA
Period 2	T	4:30:12	5:04:47	2:21:04	2:15:29	1:20:23	1:34:00	NA	NA
28	D	12.8	23.3	26.5	34.1	53.5	40.9	23.0	NA
	D*	11.3	20.1	22.7	30.6	43.4	36.3	17.7	NA
	D**	7.1	16.5	15.6	26.3	40.9	33.7	13.1	NA
Period 1	T	8:10:56	5:58:37	10:12:20	9:54:40	7:00:15	11:18:17	6:22:54	NA

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
28	D	6.4	64.5	13.1	23.8	11.1	20.4	NA	NA
	D*	5.1	59.5	10.4	19.1	9.5	18.6	NA	NA
	D**	3.4	53.4	6.1	12.9	8.6	11.1	NA	NA
Period 2	T	1:18:00	16:06:50	4:27:01	8:37:26	5:34:05	8:10:49	NA	NA
29	D	26.6	9.1	65.7	65.3	65.6	25.9	13.0	14.6
	D*	22.1	9.0	64.6	64.0	64.9	25.1	11.8	13.0
	D**	9.7	7.7	59.6	59.3	64.5	20.8	9.1	9.8
Period 1	T	10:47:58	12:40:29	13:10:29	12:22:30	2:23:20	9:19:35	6:38:41	3:55:58
29	D	0.1	18.9	10.7	26.4	10.2	17.1	23.7	72.8
	D*	0.1	18.4	10.3	24.4	9.4	16.2	22.8	72.0
	D**	0.1	10.9	8.8	21.5	5.3	11.3	20.0	70.7
Period 2	T	0:18:28	12:35:41	11:34:33	5:39:04	7:47:39	8:32:31	9:51:30	8:15:36
30	D	0.2	4.8	7.8	43.8	37.6	67.8	94.2	54.9
	D*	0.2	4.2	7.2	43.0	36.5	67.7	94.0	50.6
	D**	0.1	3.0	3.9	38.9	31.7	66.1	91.2	48.9
Period 1	T	0:15:43	6:48:17	9:51:30	10:46:58	12:43:53	11:50:47	10:59:45	7:44:59
30	D	0.1	39.2	7.9	1.8	28.5	52.1	50.1	5.0
	D*	0.1	36.6	7.6	1.7	27.7	50.5	46.2	5.0
	D**	0.0	33.1	4.6	1.3	23.9	43.9	44.0	4.7
Period 2	T	2:51:38	14:17:24	8:21:00	5:00:18	14:06:41	11:12:17	11:34:27	1:32:25
31	D	11.0	9.6	9.9	27.2	58.4	28.1	NA	NA
	D*	10.8	9.4	9.7	26.0	58.3	25.1	NA	NA
	D**	6.7	1.8	5.8	19.7	52.6	12.3	NA	NA
Period 1	T	11:43:03	14:19:32	13:22:02	13:27:18	13:15:24	10:27:29	NA	NA
31	D	15.7	58.4	17.9	52.7	24.2	17.7	17.7	NA
	D*	14.5	58.0	16.6	51.4	23.3	16.0	15.9	NA
	D**	11.6	53.5	6.9	45.5	13.7	7.4	7.4	NA
Period 2	T	7:26:18	11:04:08	13:55:58	14:10:17	13:00:10	12:11:31	12:11:31	NA
32	D	29.5	50.2	59.8	38.2	56.1	40.4	51.4	34.4
	D*	28.1	49.5	57.7	38.0	45.5	39.6	51.0	33.4
	D**	23.5	41.9	52.7	33.4	36.6	38.2	49.1	26.6
Period 1	T	6:57:57	15:02:18	9:45:11	13:30:07	10:19:54	8:31:47	10:53:35	6:06:51
32	D	6.8	51.9	6.6	66.7	45.8	33.5	84.0	165.2
	D*	5.2	46.1	5.1	63.0	40.0	32.0	81.7	164.6
	D**	3.9	36.9	3.5	56.0	31.3	23.3	74.1	163.2
Period 2	T	1:46:02	13:46:13	6:03:06	10:47:38	12:31:44	12:51:58	11:56:09	9:00:40
33	D	18.1	25.2	59.4	137.7	49.7	46.9	107.7	3.2
	D*	16.5	23.6	55.9	133.6	47.6	43.4	105.8	2.9
	D**	8.9	18.8	44.7	126.2	39.6	33.2	99.3	1.6
Period 1	T	9:12:52	10:40:14	11:09:57	16:38:39	11:35:23	8:06:21	12:10:42	1:02:13

Subject	Stat	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
33	D	28.7	29.0	35.2	137.1	75.6	3.9	85.8	1.3
	D*	23.9	26.7	31.1	128.4	68.9	3.8	78.6	1.0
	D**	16.3	21.4	21.1	120.0	63.9	2.8	67.4	1.0
Period 2	T	6:44:08	10:37:29	10:49:32	10:33:00	8:40:46	1:28:28	16:06:22	0:12:12
34	D	76.5	49.0	55.8	27.2	23.9	9.0	NA	NA
	D*	76.4	48.8	55.1	26.8	23.7	8.9	NA	NA
	D**	72.4	42.2	51.3	17.0	17.1	6.7	NA	NA
Period 1	T	9:06:51	15:56:40	11:05:04	13:44:09	13:59:42	4:00:26	NA	NA
34	D	4.5	34.4	25.7	35.9	362.1	8.3	59.0	NA
	D*	4.5	34.2	25.5	35.3	361.8	7.5	55.9	NA
	D**	4.4	28.8	18.8	32.8	356.9	5.9	47.5	NA
Period 2	T	2:17:32	16:12:29	12:46:47	9:33:54	14:20:23	3:51:11	10:59:31	NA
35	D	20.0	25.0	23.2	0.1	NA	NA	NA	NA
	D*	19.5	23.0	22.6	0.1	NA	NA	NA	NA
	D**	14.5	18.3	18.7	0.0	NA	NA	NA	NA
Period 1	T	14:59:23	24:00:00	24:00:00	5:49:37	NA	NA	NA	NA
35	D	24.4	100.3	445.2	63.4	11.1	37.0	3.3	NA
	D*	23.1	99.7	444.5	59.4	10.5	36.1	2.9	NA
	D**	19.5	94.3	441.7	50.7	6.1	31.5	0.6	NA
Period 2	T	10:10:44	10:50:34	13:01:39	14:04:23	10:12:54	10:18:35	2:31:38	NA

Appendix D

Area of classical 95% ellipse around entire series (km^2)

Table D.1: Area of 95% Ellipse

Subject	Time Period	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
1	1	24.2	0.7	1.4	0.4	16.5	0.4	0.1	NA
	2	5.3	23.7	0.2	0.1	8.3	0.0	NA	NA
2	1	26.9	0.8	11.9	0.0	NA	NA	NA	NA
	2	1190.2	145.2	2965.4	17.4	1.6	NA	NA	NA
3	1	0.0	2881.4	0.0	6.4	5.9	0.1	0.0	0.7
	2	0.0	2097.3	0.0	2.6	49.0	2.9	2851.5	24.8
4	1	0.3	37.1	35.6	118.6	3.8	39.5	NA	NA
	2	2.1	5.1	83.3	20.4	10.9	1.8	34.8	NA
5	1	29.8	876.1	12.7	3.8	34.5	0.0	NA	NA
	2	24.5	67.2	507.9	63.4	6.7	2.4	NA	NA
6	1	42.4	432.0	86.5	40.7	82.4	NA	NA	NA
	2	80.3	190.1	4.6	NA	NA	NA	NA	NA
7	1	0.5	2.4	14.5	1.9	0.8	7.1	4.5	NA
	2	2.5	0.5	0.2	9.9	1.6	5.0	2.2	0.5
8	1	0.0	8.1	3.0	0.0	0.0	2.9	NA	NA
	2	0.0	4.3	5.2	0.0	19.0	0.0	NA	NA
9	1	2.2	0.4	1572.3	0.0	0.0	0.0	NA	NA
	2	0.4	0.0	0.0	7.2	0.0	0.1	7.3	9.0
10	1	0.8	6.9	0.0	64.5	21.4	0.0	0.0	77.6

Subject	Time Period	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
11	1	25.8	21.2	123.5	16.7	52.9	14.9	0.1	0.0
	2	3702.0	127.9	3984.2	74.4	41.3	15.0	42.9	NA
12	1	5.4	10.4	19.8	24.6	77.4	1.1	0.1	0.0
	2	14.4	200.7	104.2	6.1	2.5	7.1	4.7	11.9
13	1	1.2	0.4	9.0	0.0	NA	NA	NA	NA
	2	0.0	18.4	10.4	17.9	0.0	1.6	NA	NA
14	1	12.7	8.3	0.1	2.6	14902.0	0.0	NA	NA
	2	14.7	8.6	0.7	0.0	1160.8	0.0	7.2	NA
15	1	2.3	0.8	3.3	2.5	2.0	3.2	NA	NA
16	1	57.1	142.0	49.9	6.2	3367.3	445.0	NA	NA
	2	22.5	5.5	70.0	3.5	4074.9	271.4	11.7	NA
17	1	2.0	1.1	0.7	1.6	0.0	NA	NA	NA
	2	0.0	1.3	3.9	1.1	3.0	NA	NA	NA
18	1	0.6	0.0	0.7	5.0	50.9	NA	NA	NA
	2	3.2	1.5	1.3	0.0	0.6	0.5	NA	NA
19	1	0.9	3.7	0.5	14.5	0.9	NA	NA	NA
	2	14.0	37.9	15.3	2.2	72.3	45.9	NA	NA
20	1	3.7	10.6	10.4	120.5	238.6	25.9	NA	NA
	2	8.5	7.7	11.3	1314.1	3374.8	151.1	NA	NA
21	1	35.6	6.1	28.2	0.7	21.0	0.0	7.7	NA
	2	47.5	5.9	0.0	0.3	12.0	0.0	0.0	NA
22	1	0.4	0.6	6.9	38.4	1.5	10.5	NA	NA
	2	0.2	2.6	4.5	6.9	64.0	0.4	0.2	NA
23	1	0.0	2.7	4.1	0.0	0.1	5.0	0.0	NA
	2	0.1	11.1	1.9	16.7	0.0	23.3	0.0	NA
24	1	27.4	0.2	0.0	0.0	6.3	31.4	0.0	NA
	2	0.1	0.2	0.0	0.1	2.7	7.8	82.5	NA
25	1	0.1	1.2	11.8	39.4	11.0	NA	NA	NA
	2	0.2	11485.5	13567.4	33.2	0.2	0.0	NA	NA
26	1	4.4	1002.7	8293.3	4.6	15.9	30.9	0.0	NA
	2	41.7	105.8	15.0	0.2	1.8	97.6	40.7	21.8
27	1	0.8	9.7	0.1	0.4	0.4	0.1	NA	NA
	2	0.3	0.4	0.0	0.4	0.2	4.0	NA	NA
28	1	0.1	0.6	0.1	13.0	0.7	27.0	0.2	NA
	2	0.1	139.0	0.0	0.1	1.0	0.5	NA	NA
29	1	0.2	0.8	148.5	138.5	117.6	7.2	0.8	2.3
	2	0.0	2.4	0.6	13.2	0.5	0.8	6.1	215.4
30	1	0.0	0.1	0.0	28.7	19.8	92.9	168.8	33.2
	2	0.0	8.4	0.0	0.0	11.5	10.0	31.5	2.1

Subject	Time Period	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
31	1	0.0	0.0	0.3	11.6	18.7	0.2	NA	NA
	2	1.6	20.4	0.0	20.9	0.1	0.2	0.2	NA
32	1	2.8	72.3	91.8	16.3	16.6	134.1	68.0	7.4
	2	0.1	19.3	0.3	6.5	1.5	6.7	75.0	1552.6
33	1	0.6	50.8	13.5	495.2	11.2	37.0	185.7	0.1
	2	0.4	1.4	1.5	73.7	148.0	0.1	80.3	0.0
34	1	378.5	75.3	124.0	8.3	10.7	1.1	NA	NA
	2	0.8	69.5	9.0	81.5	17224.2	0.2	17.6	NA
35	1	4.1	0.1	8.3	0.0	NA	NA	NA	NA
	2	11.5	200.7	12869.6	55.8	0.2	41.0	0.0	NA

Appendix E

Area of robust ($h = \lfloor 0.95 * n \rfloor$ good points), classical 95% ellipse and minimum spanning ellipse for all data in given time period (km^2)

Table E.1: Area of Robust and Classical 95% Ellipses, and Minimum Spanning Ellipse

Subject	Ellipse	1	2	3	4	5	6	7	8	9
Period 1	Robust	9.8	13.7	1031.8	104.9	2064.6	387.4	60.1	11.0	0.2
Period 1	Classical	53.1	47.4	1512.3	152.7	1463.0	497.6	77.6	4.7	0.7
Period 1	Spanning	128.2	65.1	1920.0	203.0	494.5	436.6	89.0	16.7	1378.7
Subject	Ellipse	1	2	3	4	5	6	7	8	9
Period 2	Robust	67.5	9263.4	3295.6	25.1	339.9	872.2	13.3	42.5	0.5
Period 2	Classical	45.6	8060.5	2706.4	33.8	238.4	872.2	14.7	31.0	6.4
Period 2	Spanning	29.6	6896.4	2493.9	134.6	504.9	373.2	12.7	22.7	39.0

Subject	Ellipse	10	11	12	13	14	15	16	17	18
Period 1	Robust	319.2	80.3	89.9	0.1	24189.4	10.5	3380.2	1.5	21.6
Period 1	Classical	319.2	58.6	126.2	6.4	16450.8	7.8	2880.2	2.3	16.4
Period 1	Spanning	176.4	86.7	93.6	17.4	11669.0	12.9	3262.9	5.4	46.4
Subject	Ellipse	10	11	12	13	14	15	16	17	18
Period 2	Robust	NA	2752.5	246.6	13.3	392.3	NA	4013.2	6.5	25.2
Period 2	Classical	NA	3426.3	251.5	31.7	628.0	NA	4013.2	6.0	16.6
Period 2	Spanning	NA	5382.7	213.8	55.5	657.9	NA	2500.6	6.8	16.6
Subject	Ellipse	19	20	21	22	23	24	25	26	27
Period 1	Robust	6.5	182.1	45.0	389.9	11.3	36.0	57.9	3680.4	13.5
Period 1	Classical	7.0	180.0	86.2	233.6	37.4	44.1	67.9	4622.5	13.1
Period 1	Spanning	12.2	256.8	125.8	193.9	28.0	75.2	67.6	6005.5	16.1
Subject	Ellipse	19	20	21	22	23	24	25	26	27
Period 2	Robust	39.1	32300.7	35.3	5.9	60.3	33.8	14045.3	160.1	0.4
Period 2	Classical	47.3	32300.7	29.3	7.6	60.3	43.9	11596.8	160.1	1.2
Period 2	Spanning	107.5	13109.0	93.8	154.3	44.4	56.2	15670.0	133.0	10.4
Subject	Ellipse	28	29	30	31	32	33	34	35	
Period 1	Robust	10.5	929.9	938.8	13.9	352.1	529.3	222.4	2.7	
Period 1	Classical	21.8	836.1	938.8	11.5	352.1	645.8	207.7	4.1	
Period 1	Spanning	59.4	2269.2	675.4	42.2	276.1	656.0	413.1	17.6	
Subject	Ellipse	28	29	30	31	32	33	34	35	
Period 2	Robust	339.9	47.0	144.4	320.5	1771.0	567.9	5028.8	5702.6	
Period 2	Classical	92.4	57.6	104.3	216.4	1171.0	429.9	5583.0	5948.9	
Period 2	Spanning	96.2	178.9	70.0	134.7	1346.8	689.1	9992.0	9281.2	

Appendix F

R Program: Large Noise Identification Methods

```
##Identifies mean and standard deviation of distances and
##  amplitude of acceleration for each non-overlapping window
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)

distAccel <- function(X,w) {
  numberwindows <- floor(length(X[,1])/w)
  dBar <- c();  aBar <- c();  dSD <- c();  aSD <- c()
  distances <- sqrt((diff(X[,"data.x"]))^2+(diff(X[,"data.y"]))^2)
  accelerations <- sqrt((diff(X[,"data.x"],differences=2))^2+
    (diff(X[,"data.y"],differences=2))^2)
        #time diff of 1 second used so this is acceleration
  for(i in 1:numberwindows) {
    dist <- distances[(w*(i-1)+1):(w*i)]
    accel <- accelerations[(w*(i-1)+1):(w*i)]
    dBar[i] = mean(dist)
    aBar[i] = mean(accel)
    dSD[i] = sqrt(var(dist))
    aSD[i] = sqrt(var(accel))
  }
  return(list(dBar,dSD,aBar,aSD))
}
```

```
#####################################################################
##             AVERAGE AMPLITUDE OF ACCELERATION             ##
#####################################################################

##Identifies windows with average amplitude of acceleration
##  above a given cut-off value
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)

aveAmp <- function(X,w,cv) {
  numberwindows <- floor(length(X[,1])/w)
  markedWindow <- which(distAccel(X,w)[[3]] > cv)
     #windows with average amplitude of acceleration greater than cv
  return(markedWindow)
}




#####################################################################
##               STANDARD DEVIATION OF DISTANCE              ##
#####################################################################

##Identifies windows with standard deviation of distances above a given cut-off value
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)

sdDist <- function(X,w,cv) {
  numberwindows <- floor(length(X[,1])/w)
  markedWindow <- which(distAccel(X,w)[[2]] > cv)
     #windows with standard deviation of distances greater than cv
  return(markedWindow)
}
```

```
#################################################################
##     STANDARD DEVIATION OF AMPLITUDE OF ACCELERATION      ##
#################################################################

##Identifies windows with standard deviation of
##  amplitude of acceleration above a given cut-off value
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)

sdAmpAccel <- function(X,w,cv) {
  numberwindows <- floor(length(X[,1])/w)
  markedWindow <- which(distAccel(X,w)[[4]] > cv)
     #windows with standard deviation of amplitude of acceleration greater than cv
  return(markedWindow)
}




#################################################################
##     RATIO OF STANDARD DEVIATION TO MEAN OF DISTANCES     ##
#################################################################

##Identifies windows with ratio of standard deviation to mean distances
##  above a given cut-off value
##X - data set (matrix or data frame)
##  x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)

sdmeanDist <- function(X,w,cv) {
  numberwindows <- floor(length(X[,1])/w)
  markedWindow <- which(distAccel(X,w)[[2]]/distAccel(X,w)[[1]] > cv)
       #windows with ratio of standard deviation to mean of distances greater than cv
  return(markedWindow)
}
```

```
#############################################################################
##   RATIO OF STANDARD DEVIATION TO MEAN OF AMPLITUDE OF ACCELERATION   ##
#############################################################################

##Identifies windows with ratio of standard deviation to mean
##  amplitude of acceleration above a given cut-off value
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)

sdmeanAmpAccel <- function(X,w,cv) {
  numberwindows <- floor(length(X[,1])/w)
  markedWindow <- which(distAccel(X,w)[[4]]/distAccel(X,w)[[3]] > cv)
     #windows with the ratio of standard deviation to
     #   mean of amplitude of acceleration greater than cv
  return(markedWindow)
}
```

Appendix G

R Program: Large Noise Filtering Methods

```
#############################################################################
##                        ELIMINATE HIGH VELOCITIES                        ##
#############################################################################

##Produce filtered series with high velocities eliminated
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##cv - cut-off value (numeric)

elimHighVel <- function(X,cv) {
  dx <- diff(X[,"data.x"]) #derivative of x values
  dy <- diff(X[,"data.y"]) #derivative of y values
  store <- sqrt((diff(X[,"data.x"]))^2+(diff(X[,"data.y"]))^2)
    #velocity series as time difference between points is 1 second
  highvel <- which(abs(store) > cv)
    #identify which absolute velocity values are above the cut-off value
  dx[highvel] <- 0 #set high velocity points to zero
  dy[highvel] <- 0 #set high velocity points to zero
  intx <- cumsum(c(X[1,"data.x"],dx)) #integrate x series
  inty <- cumsum(c(X[1,"data.y"],dy)) #integrate y series
  intSet <- matrix(NA,nrow=length(X[,1]),ncol=2)
  intSet[1,] <- intx
  intSet[2,] <- inty
  return(intSet)
}
```

```
##############################################################################
##                      ELIMINATE HIGH ACCELERATIONS                        ##
##############################################################################

##Produce filtered series with high accelerations eliminated
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##cv - cut-off value (numeric)

elimHighAccel <- function(X,cv) {
  dx <- diff(X["data.x"]) #derivative of x values
  dy <- diff(X["data.y"]) #derivative of y values
  ddx <- diff(dx) #second derivative of x values
  ddy <- diff(dy) #second derivative of y values
  store <- which(abs(ddx) > cv)
    #identify which absolute value of x-direction series acceleration values
    #   are above the cut-off value
  ddx[store] <- 0 #set high acceleration points to zero
  store <- which(abs(ddy)> cv)
    #identify which absolute value of y-direction series acceleration values
    #   are above the cut-off value
  ddy[store] <- 0 #set high acceleration points to zero
  intxx <- cumsum(c(dx[1],ddx)) #integrate acceleration series in x direction
  intyy <- cumsum(c(dy[1],ddy)) #integrate acceleration series in y direction
  intx <- cumsum(c(X[1,"data.x"],intxx)) #integrate velocity series in x direction
  inty <- cumsum(c(X[1,"data.y"],intyy)) #integrate velocity series in y direction
  intSet <- matrix(NA,nrow=length(X[,1]),ncol=2)
  intSet[1,] <- intx
  intSet[2,] <- inty
  return(intSet)
}
```

```
##############################################################################
##                        MULTI-LEVEL TRIMMED MEANS                         ##
##############################################################################

##Produce filtered series using trimmed means
##Level of filtering will depend on the level of large noise in the window
##X - data set (matrix or data frame)
##  - x values have column header "data.x"; y values have header "data.y"
##w - window length (integer)
##cv - cut-off value (numeric)
##trim - trim level for trimmed means (numeric vector of length 4)
##nSide - number of points included in trimmed mean on each side of point (integer)

multiTrimFilt <- function(X,w,cv,trim,nSide) {
  numW <- floor(length(X[,1])/w) #number of windows
  aveAmpAccel <- c()
  for(i in 1:numW) {
    ampAccel <- sqrt((diff(diff(X[(w*(i-1)+1):(w*i),"data.x"])))^2+
      (diff(diff(X[(w*(i-1)+1):(w*i),"data.x"])))^2)
    aveAmpAccel[i] <- mean(ampAccel) #average amplitude of acceleration for window i
  }
  flagW <- which(abs(aveAmpAccel)>cv[1]) #windows that have been flagged

  levelNoise <- c()
  for(i in 1:length(flagW)) { #assign noise levels to flagged windows
    if(aveAmpAccel[flagW[i]] > cv[1]) {
      levelNoise[i] = 1
    }
    if(aveAmpAccel[flagW[i]] > cv[2]) {
      levelNoise[i] = 2
    }
    if(aveAmpAccel[flagW[i]] > cv[3]) {
      levelNoise[i] = 3
    }
    if(aveAmpAccel[flagW[i]] > cv[4]) {
      levelNoise[i] = 4
    }
  }

  #Filter points that are in flagged windows
  datax <- c()
  datay <- c()
```

```
if(length(flagW)>0) { #filter if a large noise window has been identified
  if(flagW[1]==1){ #case 1: first flagged window is first window of series
    if(length(flagW)==1) { #cannot use previous points as there are none
      datax <- X[,"data.x"]
      datay <- X[,"data.y"]
    }
    else {
      for(k in 2:length(flagW)) {
        for(a in 1:w) {
          datax[w*(flagW[k]-1)+a] <- mean(X[(w*(flagW[k]-1)+a-nSide):
            (w*(flagW[k]-1)+a+nSide),"data.x"],trim=trim[levelNoise[k]])
          datay[w*(flagW[k]-1)+a] <- mean(X[(w*(flagW[k]-1)+a-nSide):
            (w*(flagW[k]-1)+a+nSide),"data.y"],trim=trim[levelNoise[k]])
        }
      }
    }
  }
  if(flagW[1]!=1) { #case 2: first flagged window is not first window
    for(k in 1:length(flagW)) {
      for(a in 1:w) {
        datax[w*(flagW[k]-1)+a] <- mean(X[(w*(flagW[k]-1)+a-nSide):
          (w*(flagW[k]-1)+a-1+nSide),"data.x"],trim=trim[levelNoise[k]])
        datay[w*(flagW[k]-1)+a] <- mean(X[(w*(flagW[k]-1)+a-nSide):
          (w*(flagW[k]-1)+a-1+nSide),"data.y"],trim=trim[levelNoise[k]])
      }
    }
  }
}

dataNewx <- X[,"data.x"]
dataNewy <- X[,"data.y"]
#substitute filtered values for unfiltered values
#   in windows identified as having large noise
if(length(flagW)>0) {
  if(flagW[1]==1) { #case 1: first flagged window is first window in series
    for(j in 2:length(flagW)) {
      for(i in 1:w) {
        dataNewx[w*(flagW[j]-1)+i] <- datax[w*(flagW[j]-1)+i]
        dataNewy[w*(flagW[j]-1)+i] <- datay[w*(flagW[j]-1)+i]
      }
    }
  }
```

```
    if(flagW[1]!=1) { #case 2: first flagged window is not first window in series
      for(j in 1:length(flagW)) {
        for(i in 1:w) {
          dataNewx[w*(flagW[j]-1)+i] <- datax[w*(flagW[j]-1)+i]
          dataNewy[w*(flagW[j]-1)+i] <- datay[w*(flagW[j]-1)+i]
        }
      }
    }

    #if last flagged window is last window in series, these points will not be
    #   filtered as there are no points after the window to include in trimmed mean
    if(flagW[length(flagW)]==numW) {
      dataNewx <- dataNewx[-c((w*(numW-1)+1):length(X[,1]))]
      dataNewy <- dataNewy[-c((w*(numW-1)+1):length(X[,1]))]
    }
    #if first flagged window is first window in series, there points will not be
    #   filtered as there are no points before the window to include in trimmed mean
    if(flagW[1]==1){
      dataNewx <- dataNewx[-c(1:w)]
      dataNewy <- dataNewy[-c(1:w)]
    }
  }
}
```

Appendix H

R Program: Time-dependent clustering algorithm

```
##med.dist.cent - maximum distance allowable between cluster centers without merging
##X - data set (matrix or data frame)
##  - x values have header "data.x"; y values have header "data.y"
##win.size - number of time points to include in scrolling window (integer)
##r - overlap r of scrolling windows (numeric)
##dist.value - radius of circle points are to be contained in to be considered a cluster
##quant.value - quantile value used to compare distances from points to center of cluster


timeCluster <- function(med.dist.cent,X,win.size,r,dist.value,quant.value) {
  num.w = floor((length(X[,1])-win.size)/((1-r)*win.size))+1
     #determine the number of windows to be scrolled through
  med.x <- c(NA) #find x value of the center of each window
  med.y <- c(NA) #find y value of the center of each window
  for(i in 1:(num.w-1)) {
    med.x[i] <- median(X[,"data.x"][(win.size*(1-r)*(i-1)+1):
      (win.size*(1-r)*(i-1)+win.size)])
    med.y[i] <- median(X[,"data.y"][(win.size*(1-r)*(i-1)+1):
      (win.size*(1-r)*(i-1)+win.size)])
  }

  quant.dist <- c(NA) #vector to store the quantile distances for each window
  for(j in 1:(num.w-1)) {
    dist.x <- c(NA)
    dist.y <- c(NA)
    dist.tot <- c(NA)
```

```
    dist.x = X[,"data.x"][(win.size*(1-r)*(j-1)+1):
      (win.size*(1-r)*(j-1)+win.size)]-med.x[j]
    dist.y = X[,"data.y"][(win.size*(1-r)*(j-1)+1):
      (win.size*(1-r)*(j-1)+win.size)]-med.y[j]
    dist.tot = sqrt((dist.x)^2+(dist.y)^2)
       #vector of distances from each point in the window to the center
    quant.dist[j] = quantile(dist.tot,quant.value)
       #finds quant.value quantile of the distances
}

  flag.vec <- rep(0,(num.w-1))
     #vector to indicate whether the window is a cluster (0-no, 1-yes)
  store.which <- which(quant.dist < dist.value)
  flag.vec[store.which]=1
  ones.places <- which(flag.vec==1) #find which windows were indicated as clusters
  ones.s <- ones.places[1]
  ones.e <- c(NA)
  store.counter.ones <- 1
  for(b in 1:(length(ones.places)-1)) { #finds groupings of 1's
    if(ones.places[b]==(ones.places[b+1]-1)) {
    store.counter.ones = store.counter.ones
    }
    else {
      ones.e[store.counter.ones]=ones.places[b]
      ones.s[store.counter.ones+1]=ones.places[b+1]
      store.counter.ones=store.counter.ones+1
    }
  }
  ones.e[store.counter.ones]=ones.places[length(ones.places)]

  meds.x <- c(NA) #finding new center x values (joining adjacent identified windows)
  meds.y <- c(NA) #finding new center y values (joining adjacent identified windows)
  marker.points <- c()
  final.x.points <- c()
  final.y.points <- c()
  num.rep <- c()
  for(d in 1:length(ones.s)) { #finding points identifed as being in clusters
    x.point <- c(X[,"data.x"][(win.size*(1-r)*(ones.s[d]-1)+1):
      (win.size*(1-r)*(ones.e[d]+1))])
    y.point <- c(X[,"data.y"][(win.size*(1-r)*(ones.s[d]-1)+1):
      (win.size*(1-r)*(ones.e[d]+1))])
    final.x.points <- c(final.x.points,x.point)
```

```
  final.y.points <- c(final.y.points,y.point)
  num.rep[d] <- win.size*(1-r)*(ones.e[d]+1) - (win.size*(1-r)*(ones.s[d]-1)+1)
  meds.x <- c(meds.x,median(x.point))
  meds.y <- c(meds.y,median(y.point))
}
meds.x = meds.x[-1]
meds.y = meds.y[-1]

if(length(meds.x)>=1) { #set up vector and counter to label clusters
  v <- rep(NA,length(meds.x))
  s <- 1
}

#join clusters that are less than med.dist.cent away from one another the same
while(length(which(is.na(v)))>=1) {
  f <- which(is.na(v))[1]
  for(i in 1:length(meds.x)) {
    g <- which(sqrt((meds.x-meds.x[f])^2 + (meds.y-meds.y[f])^2)<med.dist.cent)
    v[g]=s
  }
  s<-s+1
}

x.meds <- c()
y.meds <- c()
xy.cov <- matrix(NA,nrow=2,ncol=(2*max(v)))
for(b in 1:length(v)) {
  marker.points <- c(marker.points,rep(v[b],num.rep[b]))
}
for(ab in 1:max(v)) {
  store <- which(marker.points==ab)
  x.meds[ab] <- median(final.x.points[store])
  y.meds[ab] <- median(final.y.points[store])
  xy.cov[1,(2*(ab-1)+1)] = var(final.x.points[store]) # variance of x values in cluster
  xy.cov[2,(2*ab)] = var(final.y.points[store]) # variance of y values in cluster
  xy.cov[1,(2*ab)] = cov(final.x.points[store],final.y.points[store])
  xy.cov[2,(2*(ab-1)+1)] = xy.cov[1,(2*ab)] # covariance of x,y points in cluster
}

stores <- c()
#second step of clustering (find all points within dist.value of new center values)
for(i in 1:length(x.meds)) {
```

```
    dist.medx <- sqrt((X[,"data.x"]-x.meds[i])^2+(X[,"data.y"]-y.meds[i])^2)
    stores <- c(stores,which(dist.medx < dist.value))
  }
  return(stores)
}
```