CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Activity profiles of adults aged 50 - 70 years: functional data analysis
Authors	Weedle, Richard
Publication date	2019-10
Original Citation	Weedle, R. P. 2019. Activity profiles of adults aged 50 - 70 years: functional data analysis. MRes Thesis, University College Cork.
Type of publication	Masters thesis (Research)
Rights	© 2019, Richard Weedle. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-10-30 14:12:17
Item downloaded from	https://hdl.handle.net/10468/9977

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Activity Profiles of Adults Aged 50 - 70 years: Functional Data Analysis



M.Sc. (by Research) in Statistics
University College Cork
School of Mathematical Sciences

Richard Weedle
October 2019

Supervisors: Kathleen O'Sullivan / Dr. Tony Fitzgerald

Contents

List of Figures iii

List of Tables v

Abstract ix

Chapter 1 - Introduction 1
- 1.1 Context . . . 1
- 1.2 Data . . . 5
- 1.3 Aims . . . 11

Chapter 2 - Literature Review of Smoothing Techniques 13
- 2.1 Polynomial Regression . . . 14
- 2.2 Piecewise Polynomials . . . 16
- 2.3 Splines . . . 20
- 2.4 B-Spline . . . 24
- 2.5 Wavelets . . . 27
- 2.6 Summary . . . 38

Chapter 3 - Cluster Analysis 41
- 3.1 Methods . . . 41
  - 3.1.1 Similarity . . . 42
  - 3.1.2 Hierarchical . . . 43
  - 3.1.3 Non-Hierarchical . . . 45
- 3.2 Results . . . 47
  - 3.2.1 Similarity . . . 47
  - 3.2.2 Outlier Detection . . . 49
  - 3.2.3 Determining K . . . 51
  - 3.2.4 K-means . . . 53

Chapter 4 - Functional Principal Component Analysis 60
- 4.1 Methods . . . 60
- 4.2 Results . . . 66
- 4.3 Comparison . . . 80

Chapter 5 - Sensitivity Analysis 84
- 5.1 FPCA . . . 85
- 5.2 Smoothing . . . 89
- 5.3 Epoch . . . 96
- 5.4 Weekday vs weekend . . . 97

Chapter 6 - Discussion **99**
6.1 Discussion of findings 99
6.2 Limitations and Recommendations 100
6.3 Implications 102
6.4 Conclusions 103

References **105**

Appendices **112**

List of Figures

1.1 Tri-axial GeneActiv accelerometer 6
1.2 Collapsed 1 minute data for ID: 8 8
1.3 Stratified data for ID: 8 by intensity category 9
1.4 Aggregated weekday activity for ID: 8 9
2.1 Polynomials with increasing degree: (a) 1^{st}, (b) 3^{rd}, (c) 5^{th}, (d) 7^{th} 15
2.2 Piecewise polynomials with two knots and orders; (a) Zero, (b) One, (c) Two, (d) Three 17
2.3 Piecewise continuous polynomials with one knot and increasing orders. (a) One, (b) Two, (c) Three 18
2.4 Piecewise continuous linear polynomials for one knot, (a) without continuous 1^{st} derivative, (b) with continuous 1^{st} derivative 19
2.5 Cubic splines with uniformly distributed knots. (a) Two, (b) Three, (c) Four, (d) Five 21
2.6 Cubic splines with knots at times of high variability 22
2.7 Cubic splines extrapolated beyond their boundaries with uniformly distributed knots. (a) Two, (b) Three, (c) Four, (d) Five 23
2.8 Basis functions for a B-spline with order 1 25
2.9 Basis functions for a B-spline with order 2. (a) without augmented knots, (b) with augmented knots 26
2.10 Basis functions for a B-spline with orders: (a) 3, (b) 4 26
2.11 Wavelet examples (a) Haar, (b) Mexican hat, (c) Symmlet, (d) Daubechies 28
2.12 Select translations and dilations of the Haar wavelet family 29
2.13 Orthogonality example 30
2.14 Haar father wavelet 31
2.15 Wavelet decomposition tree 32
2.16 Scale approximations. (a) 1^{st}, (b) 2^{nd}, (c) 3^{rd}, (d) 4^{th} 33
2.17 Haar with details (a) 1^{st} scale approximation, (b) Raw details, (c) Thresholded details 34
2.18 Daubechies mother and father. (a) m=1, (b) m=2, (c) m=4, (d) m=8 36
2.19 Daubechies-4 DWT with increasing scale approximations. (a) 3^{rd}, (b) 4^{th}, (c) 5^{th}, (d) 6^{th} 37
2.20 6^{th} scale Daubechies - 4 and 8 WT 40
3.1 Similarity measure between IDs 8 and 138 42
3.2 Dendrogram example 44
3.3 Dendrogram example for single linkage 44
3.4 Elbow method example 45
3.5 First three clusters with agglomerative approach. (a) 2965 & 3002 (1520), (b) 3355 & 3159 (1755), (c) 2994 & 2985 (1757) 48
3.6 Hierarchical clusters formed using single linkage. (a) Full, (b) Truncated 49

3.7 Boxplot of total physical activity . . . 51
3.8 Hierarchical clusters: Complete linkage . . . 51
3.9 Complete linkage: Agglomerative coefficient vs. Number of clusters . . . 52
3.10 Centroids for K-Means clustering. (a) K=4, (b) K=5, (c) K=6 . 53
3.11 Four cluster solution. (a) 1, (b) 2, (c) 3, (d) 4 . . . 54
3.12 N-types. (a) k=5, (b) k=6 . . . 56
3.13 Sum of squared distances vs K . . . 56
3.14 Silhouette analysis. (a) K=5 (b) K=6 . . . 57
3.15 Lowest silhouette score (a) 1: N-type (High), (b) 2: E-type, (c) 3: M-type, (d) 4: N-type (Moderate) , (e) 5: N-type (Low) . . . 58
4.1 Profiles for IDs 1 and 8 (a) Normal, (b) Mean centred . . . 62
4.2 Scatterplot of X_1 vs X_2 with eigenvectors . . . 63
4.3 FPCs (a) ϕ_1, (b) ϕ_2, (c) ϕ_3, (d) ϕ_4 . . . 66
4.4 Cumulative explained variance vs. No. of FPCs . . . 67
4.5 FPCA recomposition for ID 8. (a) k=1, (b) k=3, (c) k=7, (d) k=11 (e) k=18 . . . 69
4.6 Histograms of FPC scores. (a) c_{i1}, (b) c_{i2}, (c) c_{i3}, (d) c_{i5}, (e) c_{i7} . 71
4.7 Mean curve +/- 1 standard deviation. (a) ϕ_1, (b) ϕ_3, (c) ϕ_7, (d) ϕ_{11} 73
4.8 c_{i1} vs. c_{i2} . . . 74
4.9 FPCA recomposition for ID 2874. (a) μ, (b) $k = 1$, (c) $k = 2$ . . 74
4.10 FPCA signal recomposition for ID 3539. (a) μ, (b) $k = 1$, (c) $k = 2$ 75
4.11 First FPC (ϕ_1) scores. (a) N-type: High, (b) N-type: Low . . . 76
4.12 FPC scores for ID 2874 and 3539 . . . 76
4.13 Centroid reconstruction from average FPCA scores . . . 77
4.14 Average FPCA scores (a) All clusters, (b) High (1), Moderate (4) & Low (5), (c) Evening(2) & Morning (3) . . . 78
4.15 Box plots for the first FPC scores for each cluster . . . 78
4.16 FPC comparisons (a) All clusters: ϕ_1 vs. ϕ_2, (b) High (1), Moderate (4) & Low (5): ϕ_1 vs. ϕ_2, (c) Evening(2) & Morning (3): ϕ_2 vs. ϕ_3 . . . 79
4.17 Clustering results (a) Distance, (b) FPCA . . . 80
4.18 Distance between centroids (a) High (1) vs. Evening (2), (b) High (1) vs. Morning (3), (c) Evening (2) vs. Morning (3) . . . 82
5.1 Cluster centroids (a) 18 components, (b) 3 components . . . 85
5.2 14^{th} FPC . . . 87
5.3 Reconstruction with 13 and 14 FPCs. (a) 2356, (b) 2676 . . . 88
5.4 Curve reconstruction for ID 2356. (a) 13 FPCs, (b) 14 FPCs . . . 89
5.5 Cluster centroids (a) DB-4, (b) Cubic spline - 23 knots, (c) Cubic spline - 11 knots . . . 91
5.6 Cluster constituent changes (a) ID:1143, DB4, (b) ID:1143, CS23, (c) ID:2294, DB4, (d) ID:2294, CS23 . . . 93
5.7 ID:1143 scale approximations. (a) 6^{th}, (b) 5^{th}, (c) 4^{th} . . . 94

5.8 ID:1143 scale approximations for 5 minute epoch. (a) 3^{th}, (b) 4^{th}, (c) 5^{th} 96
5.9 Profiles for ID 1143 using data from the weekdays and full week . 97
5.10 Cluster centroids. (a) Weekday only, (b) Full week 98
A.1 Age distribution (a) Accelerometer group, (b) Full cohort 112
A.2 BMI distribution (a) Accelerometer group, (b) Full cohort 112
A.3 Education distribution (a) Accelerometer group, (b) Full cohort . 113
A.4 Smoking status distribution (a) Accelerometer group, (b) Full cohort 113

List of Tables

1.1 MET intensity classifications 3
1.2 Examples of sedentary, light, moderate and vigourous activities . 3
1.3 Mitchelstown cohort 5
1.4 Raw accelerometer data 7
1.5 Accelerometer cut points 8
2.1 MSE for smoothing techniques 39
3.1 Similarity measure: Euclidean distance 48
3.2 Potential outliers 50
3.3 Top ten individuals by total activity 50
3.4 Complete linkage: Agglomerative schedule 52
3.5 Silhouette mean distances for an individual to all other clusters . 59
4.1 Variable means 61
4.2 Functional principal components explained variance 67
4.3 First 10 principal component weights for first 10 IDs 70
4.4 FPC scores: Standard deviation and absolute mean 72
4.5 K-means on FPC scores 77
4.6 Cluster agreement: FPCA vs. Distance method 80
5.1 Cluster agreement: 18 vs. 3 FPCs 86
5.2 Cluster agreement: 18 vs. 17 FPCs 86
5.3 Concordance percentages: 17 to 3 FPCs 87
5.4 Cluster agreement: 18 vs. 13 FPCs 87
5.5 Cluster change IDs: 18 vs 13 FPCs 87
5.6 Cluster agreement: DWT - DB4 vs. Cubic spline (23 knots) . . . 92
5.7 Cluster change IDs: DWT - DB4 vs. Cubic spline (23 knots) . . . 92
5.8 Cluster agreement: DWT - DB4 vs. DWT - DB8 94
5.9 Concordance percentages: 5^{th} and 4^{th} scale approximations 95
5.10 Concordance percentages: 5 minute epoch with 3^{rd}, 4^{th} and 5^{th} scale approximations 96
5.11 Cluster change IDs: 1 minute vs. 5 minute epochs 97

Acronyms

PA Physical Activity

VPA Vigorous Physical Activity

MVPA Moderate to Vigorous Physical Activity

WHO World Health Organisation

PCA Principal Component Analysis

FPCA Functional Principal Component Analysis

FPC Functional Principal Component

DWT Discrete Wavelet Transform

DB Daubechies

CS Cubic Spline

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

_____

Richard Weedle

_____

Date

Acknowledgements

I would like to thank my supervisors Kathleen O'Sullivan and Tony Fitzgerald for their support and guidance throughout the course of my study. Their constructive comments, critiques and suggestions throughout each iteration of this thesis helped me remain focused throughout.

They suggested a topic for this thesis and then allowed me to go through the research process in my own way. This permitted me to gain a deeper understanding of statistics and ensuring that I truly benefitted from this M.Sc. programme.

Abstract

Physical activity has a major impact on health. Questionnaires are the most common method of physical activity assessment. While cost effective, these are subjective and can correlate poorly with actual activity levels. Accelerometers have gained popularity given their accuracy, objectivity and ability to capture large amounts of data. Simple summary measures such as the total or average activity over the day are often used. However, these fail to exploit the longitudinal nature of the data and do not capture the variation in activity levels throughout the day. This study intends to capitalise on this nature by implementing a functional data analysis approach.

Activity data was collected from a cohort of 475 people in Mitchelstown in 2011. The individuals wore wrist worn accelerometers in a free living environment for a week. This data was collapsed into 1 minute epochs and each epoch was then aggregated over the week to get an estimate of daily circadian activity. The discrete wavelet transform was chosen as the smoothing technique to reveal the underlying functional nature of the data. This allows every individual in the cohort to be represented by a smooth activity profile. This study aimed to identify and characterise subgroups within a cohort based on these activity profiles.

Functional principal component analysis was applied to these activity profiles in order to explore the dominant patterns within the data. Each individual's profile was approximated by a weighted sum of profiles and these weights were then used to perform a cluster analysis. Five distinct subgroups were identified. These differed from each other in both the magnitude of the activity and the times at which the activity occured. A more simplified approach, based purely on the distance between profiles, was also implemented. Two distinct clustering methods identified the exact same 5 subgroups in the cohort. To ensure their robustness, these results were subject to a sensitivity analysis with respect to the epoch length, smoothing technique and number of functional components utilised in the clustering.

Other studies have clustered accelerometer data in terms of absolute activity volume, as in high or low activity groups. However, they do not place too much value in using the granularity of the data to determine what time of day people are active. In addition to the high, moderate and low activity subgroups, our analysis revealed two subgroups which have a propensity to be active in either the morning or evening. It is suggested that these are indicative of an individual's biological rhythm or chronotype. The Mitchelstown cohort was re-screened 5 years later in 2016, which presents an exciting opportunity to examine changes in these profiles over time.

Chapter 1 - Introduction

The first chapter of this thesis opens with some context for physical activity (PA). It describes what PA is and outlines the potential detriments of inactivity. Different methods of assessing PA and the minimum activity requirements for health benefits are discussed. This is followed by an introduction to the data that will be used in this study before concluding with a synopsis of what to expect in subsequent chapters.

1.1 Context

PA is defined as any bodily movements produced by skeletal muscles that results in energy expenditure greater than at rest and which is health enhancing (Caspersen, Powell, & Christenson, 1985; Waxman, 2004). This is a broad definition and covers all types of activity, including walking, cycling, gardening, housework, sport, dancing and anything else that requires movement. PA consists of the following dimensions: frequency (how often the activity occurs); intensity (how strenuous the activity is); time (how long the activity lasts), and type (the actual activity type) (Pate et al., 1995).

PA has a major impact on health. Regular PA is the key to getting healthy and staying healthy. It is recognised that PA is a major independent modifiable risk factor for chronic diseases, such as coronary heart disease (CHD), type 2 diabetes, stroke, cancer, osteoporosis and depression (D. of Health & Children, 2009; Pate et al., 1995; Yusuf et al., 2004; Eckel, Krauss, et al., 1998). However, studies (Morgan et al., 2009; Ipsos et al., 2016) show that few Irish people take part in regular physical activity.

Why are some people more physically active than others? The health benefits of exercise and being active are clear (D. of Health & Children, 2009; Pate et al., 1995; Yusuf et al., 2004; Eckel et al., 1998). Understanding the factors that influence physical activity can aid the design of more effective targeted interventions (Heath et al., 2012). In these interventions it would be valuable to know whether characteristic patterns of physical activity are associated with particular population subgroups.

The unequivocal link between physical activity and health has prompted researchers and public health officials to search for valid, reliable, and logistically feasible tools to measure and quantify free-living physical activity. There is a need for assessing the prevalence of physical activity engagement, identifying active and inactive segments of the population, and evaluating the effectiveness of interventions.

The National Guidelines on Physical Activity for Ireland (D. of Health &

Children, 2009) adopted the World Health Organisation's (WHO) global recommendations on physical activity for health. These guidelines emphasise the importance of PA and outline the recommendations for PA for people of all ages. They establish a national consensus for appropriate levels of PA to enhance health. "The key message is that physical activity is for everyone, and any level of activity is better for your health than none" (WHO).

Specifically it states that adults aged 18 - 64 years, should engage in moderate active for at least 30 minutes a day on any 5 days of a week (or 150 minutes per week). To follow this guideline, an interpretation as to what constitutes moderate activity is required. Broadly speaking it can be thought of as activity that increases your breathing and heart rate, but you are still able to maintain a conversation.

While increasing PA benefits everyone in terms of health, there is strong evidence that the greatest benefits occur when the least active become moderately active (Nocon et al., 2008). The Healthy Ireland survey (Ipsos et al., 2016) found that 65% of people in Ireland were aware of these guidelines. It found that 56% believed they undertook a sufficient level of activity and that 32% actually did. This highlights the need for an accurate and objective measure of PA. The survey also found that 91% of people who felt that they do not undertake a sufficient level of activity would like to be more physically active. This means that 9% are happy with a sedentary lifestyle and are willing to accept the health risks of inactivity.

There is no gold standard for measuring PA (Welk, 2002), as no single instrument is able to record cardioresporatory characterisitcs and behavioural response during PA. Activity can be measured in metabolic equivalents, or METs (Ainsworth et al., 2000). METs relates to the rate of the body's oxygen uptake for a given activity as a multiple of the resting volume of oxygen consumption. One MET is defined as the amount of oxygen consumed while sitting at rest. It is the ratio of work metabolic rate to a resting (basal) metabolic rate. Basal metabolic rate (BMR) refers to the number of calories a body burns each day to stay alive. It is the energy required by someone to perform the basic functions like breathing and the circulating of blood. BMR does not include physical activity, the process of digestion, or things like walking from one room to another. It is the number of calories someone would expend in a 24 hour period if all they did were lie in bed all day long.

Every activity has a MET value which calculates the energy required for that activity. One MET is the energy expended while at rest, like sitting quietly or sleeping. Specifically sedentary behaviour refers to any waking activity characterised by an energy expenditure $<$1.5 METs (Bames et al., 2012). This usually refers to any time someone is sitting or lying down, such as watching

TV, driving, computer use or reading. A two MET activity expends twice the amount of energy per minute than at rest. If a person does a two MET activity for 30 minutes, he/she has done 60 MET-minutes. Pate et al. (1995) proposed a model for classifying the MET intensity of physical activities, which is still used in today (Ainsworth et al., 2011; D. of Health & Children, 2009). This is shown in Table 1.1.

Intensity	METs
Sedentary	<1.5
Light	1.5 - 3
Moderate	3-6
Vigorous	>6

Table 1.1: MET intensity classifications

Knowing these cut points allows different activities to be quantified in terms of their MET equivalents. Examples of activities for each of the sub groups sedentary, light, moderate and vigourous are presented in Table 1.2.

Sedentary	Light	Moderate	Vigourous
Sitting	Slow walk	Brisk walk	Jogging
Watching TV	Cooking	Mowing the lawn	Shovelling
Driving	Washing the dishes	Light bicycling	Fast bicycling
Computer use	Playing most instruments	Tennis doubles	Tennis singles

Table 1.2: Examples of sedentary, light, moderate and vigourous activities

Moderate activity is defined as 3 - 6 METs, they are activities that generate enough movement to burn off 3 to 6 times as much energy per minute than while at rest.

To get accurate measurements of METs, it would be necessary to measure an individual's oxygen consumption using a portable metabolic system. These are not readily available, therefore simpler, alternative methods such as questionnaires or accelerometers are used to give an estimate or proxy.

While a lack of PA is a prominent risk factor for diseases, research however is often hindered by the challenge of employing a valid, reliable measure that addresses the research question (Sylvia, Bernstein, Hubbard, Keating, & Anderson, 2014). Questionnaires are the most common method of PA assessment (Castillo-Retamal & Hinckson, 2011; Lagerros & Lagiou, 2007). They are cost effective, easy to administer and are useful for determining discrete categories of activity level (e.g., low, moderate, high). They are, however, subjective and less robust in measuring light or moderate activity (Jacobs, Ainsworth, Hartman, & Leon, 1993). They are also known to correlate

poorly with actual activity levels, often overestimating them (Matthews & Freedson, 1995; Coleman, Saelens, Wiedrich-Smith, Finn, & Epstein, 1997). This means that it is difficult to assess whether or not an individual meets the recommend guidelines for activity.

One example of a PA questionnaires, is the International Physical Activity Questionnaire (IPAQ) (Hagströmer, Oja, & Sjöström, 2006), which is a self-reported questionnaire that asks you about your activity in the last 7 days. It measures the duration and frequency of PA in the following domains:

- Job-related
- Transportation
- Housework, house maintenance, caring for family
- Leisure time, recreation and sport
- Time spent sitting

Minutes spent in each activity are multiplied by the MET equivalent and summed, in order to calculate an individual's MET-minutes. If this is within the 500 to 1000 range guideline, then the person is deemed to be active enough to have met the PA guidelines. These MET equivalents are sourced from the Compendium of Physical Activity (Ainsworth et al., 2000), which provides a full list of activities and their MET equivalents. This is a coding scheme that classifies specific PA by rate of energy expenditure. The guidelines stated previously, 30 minutes per day for any 5 days in a week, have an equivalent MET-minutes target of 500 to 1000. Total weekly activity should be in the range of 500 to 1000 MET-minutes of moderate to vigourous activity to produce substantial health benefits (U. D. of Health, Services, et al., 2018).

Another method for PA assessment are accelerometers, which have gained in popularity given their accuracy, objectivity and ability to capture large amounts of data. They are motion sensors that detect accelerations produced by the human body. Given that acceleration is defined as the rate of change in velocity, the frequency, intensity and duration of PA can be assessed through body movement. Within an accelerometer are transmitters that are stressed by acceleration forces, which leads to an electrical signal being produced that is converted to provide an indication of movement (Welk, 2002).

Tramsmittors measure acceleration in real time and can detect movement in up to three orthogonal planes. Devices can be worn in numerous places on the body, including the wrist and hip. These different body placement positions result in different signal patterns and accuracies (Kangas, Konttila, Winblad, & Jamsa, 2007), further complicating comparisons between studies. It is been

found that PA estimates can vary by as much as 41% across wear locations (Kerr et al., 2017). Asking subjects to wear the accelerometers on the wrist instead of the hip leads to increased wear time (Mannini, Intille, Rosenberger, Sabatini, & Haskell, 2013). How to calculate the individual's wear time will be discussed in section 1.2.

Current research-grade accelerometers allow investigators to apply various methods to convert raw acceleration data to PA metrics, such as time spent in various intensity categories (Matthew, 2005). Equivalent cut off points, specific to the brand of accelerometer, can be used to group PA into sedentary, light, moderate and vigorous sub bands. This is explained further in section 1.2 along with an introduction to the data that will be used in this study, which includes accelerometer data.

1.2 Data

The original Cork and Kerry Diabetes and Heart Disease Study - Phase 1, was undertaken in 1998, and the cohort was recruited from across 17 different general practices in Cork and Kerry (Perry et al., 2002). Phase 2 began in 2008 and a new cohort of 2047 men and women, aged 50 to 69 years, was recruited from a single large primary care centre, the Livinghealth Clinic, in Mitchelstown (Kearney, Harrington, Mc Carthy, Fitzgerald, & Perry, 2012). This primary care centre includes 8 general practitioners and serves a catchment area of approximately 20,000 with a mix of urban and rural residents. At baseline, this new cohort completed both a questionnaire and physical assessment during the study period which ran between April 2010 and May 2011. This study will focus on data from this Mitchelstown cohort, a breakdown of which can be found in Table 1.3.

Demographic characteristics	Men (%)	Women (%)	Total (%)
Age			
50-54 years	249 (25.1)	261 (25.6)	510 (25.4)
55-59 years	285 (28.8)	272 (26.7)	557 (27.7)
60-64 years	260 (26.2)	289 (28.4)	549 (27.3)
65-69 years	197 (19.9)	196 (19.2)	393 (19.6)
Total	991	1018	2009

Table 1.3: Mitchelstown cohort

As can be seen from Table 1.3, 2009 individuals are retained for analysis rather than the original 2047. This is due to the exclusion of 2 individuals under 50 years of age, and 36 over the age of 70. Table 1.3 illustrates that the ages for this cohort are close to uniformly distributed between 50 and 70. In addition to age, marital status, education and other participant characteristics were also recorded. Among others, these included:

- Body Mass Index (BMI) - A person's weight (kg) divided by their height in metres squared. For the classifications, normal is less than 25 kg/m^2, overweight between 25 and 30 kg/m^2, and obese is greater than 30 kg/m^2 (Organization, 2000).

- Smoking status - 3 classifications were used; never smoked, former smoker and current smoker.

- Psychological well being - The Center for Epidemiologic Studies Depression (CES-D) scale (Radloff, 1977) was used to determine whether a person was depressed and to what extent.

An objective measurement of PA was introduced into the study in January 2011. A subsample of the participants (745) in the Mitchelstown cohort were asked to wear a tri-axial GeneActiv accelerometer, as shown in Figure 1.1, on their wrist in a free-living environment for a week.



Figure 1.1: Tri-axial GeneActiv accelerometer

Of the 745 who were asked, 475 agreed to wear the accelerometer (44.6% males, mean aged 59.6 years (SD=5.5)). Only 745 people out of the full cohort were asked as the accelerometers were introduced late in the study. The accelerometers are waterproof and can be worn 24 hours a day. It was set to record at 100Hz, or 100 readings per second. Each measurement gave the acceleration along the x, y and z axis. A snapshot of this raw data is shown in Table 1.4. The table shows ten readings, corresponding to a tenth of a second.

Time	x	y	z
10:25:03:600	0.79	0.23	-0.42
10:25:03:610	1.13	0.28	-0.35
10:25:03:620	1.91	0.29	-0.34
10:25:03:630	3.1	0.12	-0.44
10:25:03:640	3.96	-0.34	-0.53
10:25:03:650	4.1	-0.9	-0.62
10:25:03:660	4.02	-1.49	-0.58
10:25:03:670	3.94	-1.89	-0.62
10:25:03:680	3.79	-1.99	-0.75
10:25:03:690	3.36	-1.94	-0.65

Table 1.4: Raw accelerometer data

As part of the data processing, wear and non-wear time needed to be determined. If a participant had less than 10 hours of wear time activity on any given day, they were excluded from the study. This was done using a procedure identified by Van Hees et al. (2011). Non-wear time was calculated for each accelerometer axis on the basis of the standard deviation and the value range, for successive 30 minute blocks. If the standard deviation was below a certain threshold, the block was categorised as non-wear. From the 475 who agreed, 397 had valid accelerometer data.

The Euclidean norm of the acceleration in x, y, and z axes was calculated to turn this raw data into a metric for activity. In order to separate out the activity related component of the acceleration signal, one gravitational unit was subtracted from the vector magnitude. This produced a gravity adjusted signal magnitude vector (SVM_{gs}), which will be used as our measure of activity. This calculation is shown in Equation (1.1).

$$SVMgs = \sum (\sqrt{x^2 + y^2 + z^2}) - 1 \tag{1.1}$$

where x, y and z are in units of gravity. The reason for subtracting 1 here, is that when the accelerometer is static and the earth's gravitational pull is the only acceleration, this result will be zero. SVM_{gs} is calculated for every measurement. For each individual, 100 measurements per second for a week, results in approximately 60 million measurements. To make this more manageable, these SVM_{gs} calculations are collapsed over a specific time interval or epoch. In this study, 1 minute epochs will be considered. For a single individual this corresponds to 1440 measurements per day, or 10080 for a week. To give an idea of these activity patterns, the first individual in the cohort (ID: 8) is chosen as an example. The collapsed 1 minute activity epochs is plotted against the day of the week, and is shown in Figure 1.2. The labels on the x-axis are at the midday point for each day.

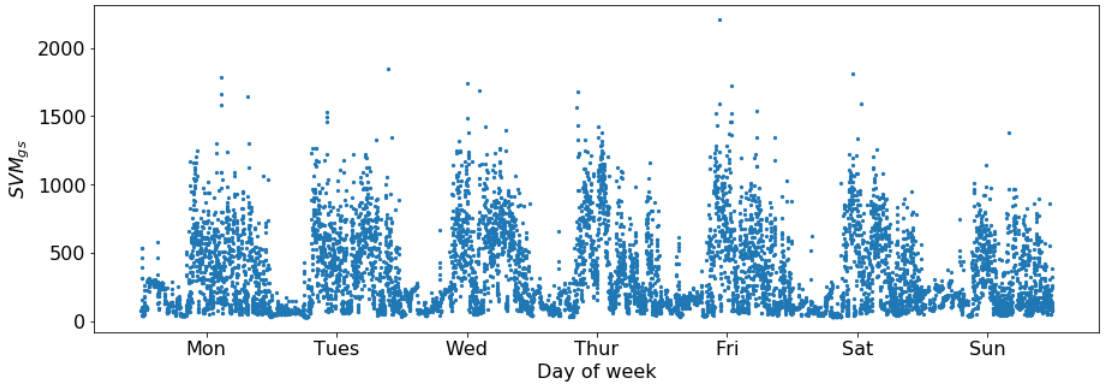


Figure 1.2: Collapsed 1 minute data for ID: 8

The cyclical pattern seen in Figure 1.2 corresponds to the individual's sleep/wake cycle. They are active during the day and relatively inactive in the evenings while they are asleep. Each epoch can be categorised according to intensity, based on validated cut off points, in terms of SVM_{gs}, for this particular brand of accelerometer (Dillon et al., 2016). These labels are the same as for METs, which are sedentary, light, moderate and vigorous. The cut off points are given in Table 1.5.

Intensity	Cut points (SVM_{gs})
Sedentary	$<$700
Light	700 - 1087
Moderate	1088 - 2180
Vigorous	$>$2180

Table 1.5: Accelerometer cut points

If these limits are applied to the example individual (ID: 8), Figure 1.2, the time spent in the different intensity categories can be visualised. This is shown in Figure 1.3. Essentially, a cut point applied to the data assumes that an epoch that scores higher than this value is indicative that the individual has been vigorously active, for example, for the duration of that epoch.

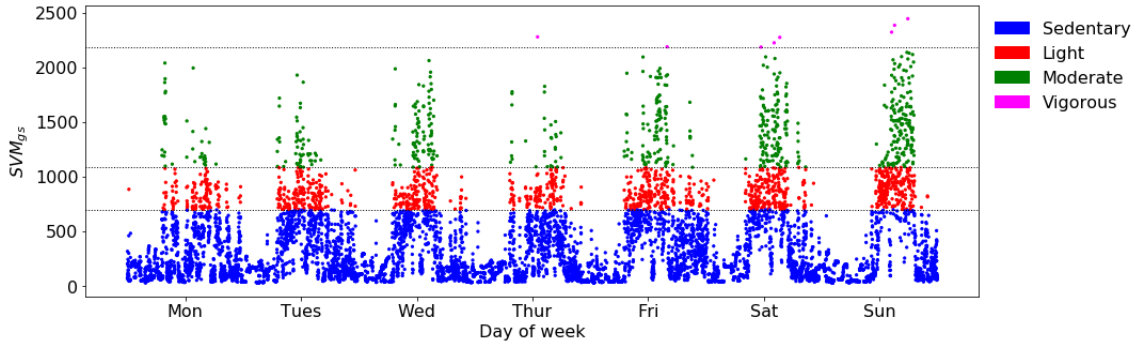


Figure 1.3: Stratified data for ID: 8 by intensity category

This particular individual spent the majority of time (8169 minutes) in the sedentary category and 8 minutes in the vigorous intensity category over the course of the week. They also spent 1293 minutes in the light category and 610 in moderate. Using the guidelines of 500 - 1000 MET minutes of moderate to vigorous activity, then this individual has met the guidelines for this week. Measures such as these fail to take advantage of the longitudinal nature of the data.

Rather than trying to emulate what other studies (Trost, Kerr, Ward, & Pate, 2001; Tucker, Welk, & Beyler, 2011; Troiano et al., 2008) have done by trying to convert the accelerometer readings into their MET equivalents, this study will attempt to leverage the longitudinal nature which is neglected in these summary methods. To get an estimate for the daily circadian activity profile, each epoch was then averaged over 5 days, Monday to Friday. The weekend data was not included to avoid introducing variation between weekday and weekend activity patterns. For this cohort, it was previously discovered that sedentary and light activity differ on Sunday compared to the rest of the week (Dillon et al., 2016). The PA profile for the example individual, ID: 8, in 1 minute epochs is shown in Figure 1.4.

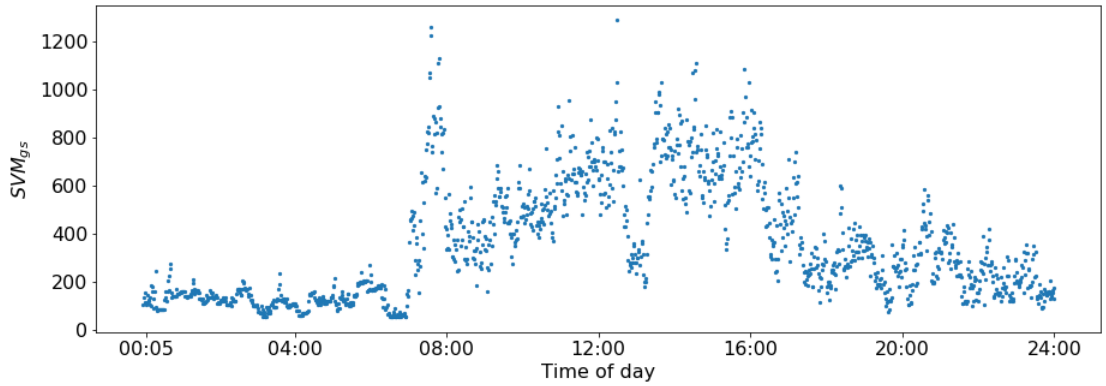


Figure 1.4: Aggregated weekday activity for ID: 8

The period from midnight to 8 a.m. has consistently low PA in this example and so could be interpreted as this individual being asleep. There are 1440 data points for this individual, and the resulting daily activity profiles are irregular functional data characterised by many peaks representing short bursts of intense activity. As self-reported measures have been shown to be unreliable (Washburn & Montoye, 1986), systems for objective activity profiling can play an important role in epidemiological studies. These systems can also be used to assess the effectiveness of different interventions aimed at increasing physical activity in individuals.

Some current approaches to analyse data such as this are based on simple summaries. For example, 30 minute averages (Cradock et al., 2004), average daily activity level (Talbot, Gaines, Huynh, & Metter, 2003), or the proportion of time spent above specific cut off levels that correspond to low, moderate and vigorous activity (Abbott & Davies, 2004). These summaries have their limitations as they do not make full use of the rich information contained in the functional data. They do not account for time of day variability and conclusions drawn from the arbitrary choice of 30 minute intervals may be sensitive to the choice of endpoints for these intervals.

In a study by Lee, Yu, McDowell, Leung, & Lam (2013), accelerometer data was collected during 2009 to 2011 for 1714 participants in Hong Kong. Two clusters were identified, one more active than the other. The active had a routine PA pattern on weekdays and a more varied pattern on weekends. The less active cluster had consistently low PA patterns on both weekdays and weekends. The conclusions of the study suggest that potential interventions to promote PA would be most effective in targeting those who are sedentary at weekends, suggesting free weekend PA programmes.

A study by Staudenmayer, Pober, Crouter, Bassett, & Freedson (2009) used cluster analysis to identify physical activity type from accelerometer data. Clustering the accelerometer signals determined four categories: 1) very low mean signals (low level of activity); 2) rhythmic and repeatable signals (locomotion); 3) less rhythmic and lower mean signals (household activities/other); and 4) high variability and high mean signals (vigorous sports). Based on a reading of the literature, it is foreseeable that subgroups will exist within the cohort.

1.3 Aims

The aim of this study is to identify subgroups in the cohort based solely on their activity profiles. An effort will be made to then characterise the individuals in each group.

Methods that model these activity profiles in their entirety have the possibility of extracting more information than summary approaches. Functional data analysis (FDA) is a general name for approaches that consider the functional profiles rather than a collection of data points (Ramsay & Silverman, 2007). The irregularity of accelerometer profiles, as seen in Figure 1.4, makes modelling in this way challenging. Data smoothing can be used to allow important patterns to stand out and reveal the underlying functional nature of the data. Many different smoothing techniques are employed in studies involving functional data analysis (Ullah & Finch, 2013). A subset of these techniques will be reviewed in Chapter 2.

Selected methods from this discussion will then be applied to the data from the Mitchelstown Cohort, who wore the accelerometers, in order to determine the optimal smoothing solution. They will be used to explore patterns in the activity level profiles. Chapter 3 will then explore clustering methods in order to identify subgroups or sub-profiles within the cohort based on these activity profiles. Clustering is a classification technique and its goal is to discover the natural groupings of a set of patterns, or in this case, profiles. Those within each cluster are more closely related to one another than those assigned to different clusters.

Chapter 4 will explore the dominant patterns in the data through Functional Principal Component Analysis (FPCA). The output from this analysis will again be used to perform cluster analysis. The results of this will be compared and contrasted to the previous method.

Chapter 5 will investigate the benefits and drawbacks to the way the data has been collapsed and aggregated. Can you tell activity patterns averaging over 5 days or are they lost? Would the groupings be different? The statistical techniques used will also be analysed. The sensitivity to the number of principal components used will be considered, as well as the choice of smoothing technique prior to the derivation of the principal components.

Chapter 6 will then provide an overall discussion. It will outline the findings, their context and limitations before providing the implications and conclusions.

The data manipulation and analysis performed throughout this thesis was implemented mostly using the Python programming language (Rossum, 1995).

The following Python libraries were utilised:

- Scikit-learn (Pedregosa et al., 2011). This package consists of tools for data mining and analysis, as well as algorithms for classification, regression and clustering.
- Scipy (Jones, Oliphant, & Peterson, 2014). Used for computing such things as linear algebra, interpolation and optimization.
- Numpy (Van Der Walt, Colbert, & Varoquaux, 2011). Adds support for manipulation of large, multi-dimensional arrays and matrices, and functions to operate on these arrays.
- Pandas (McKinney et al., 2010). Offers data structures and operations for manipulating numerical table structures and time series.
- Matplotlib (Hunter, 2007). Library used for plotting.

In addition to Python, R (Team et al., 2013) was used for the implementation of functional principal component analysis. This analysis utilised the R package fdapace (Dai, Hadjipantelis, Ji, Mueller, & Wang, 2017).

Chapter 2 - Literature Review of Smoothing Techniques

Data analysis can be broadly classified into two types (Tukey, 1977):

1. Exploratory/Descriptive - The investigator does not have pre-specified models or hypotheses but wants to understand the general characteristics or structure of the data.

2. Confirmatory/Inferential - The investigator wants to confirm the validity of a model/hypothesis given the available data.

The first type is the focus of this chapter, where a number of smoothing techniques will be discussed followed by the application of some cluster analysis.

One salient feature of functional data is that, although the underlying functions are often continuous and smooth, data can only be collected discretely, which often produces measurement errors. Therefore, smoothing is often the first step in any Functional Data Analysis (FDA), and its purpose is to convert raw discrete data points into a smoothly varying function (Ullah & Finch, 2013). Smoothness at its most simple, means there are no corners, or in mathematical terms the smoothness of a function is a property measured by the number of derivatives it has that are continuous. It is useful to emphasize any underlying patterns which may be evident in the data. The choice of smoothing technique is dependent upon the underlying behaviour of the data being analysed (Ramsay, 2005).

Smoothing can reduce the dimensionality of the data. This is often a precursor for clustering techniques, as applying these techniques to the raw observations does not utilise the underlying functional structure of the data. Given a set of data-points, such as those in Figure 1.4, a smooth curve that approximates the points is determined. Smoothing algorithms should be efficient and not overly sensitive to round-off errors in the computations (Lyche & Morken, 2008). Smoothing can also be viewed as a time series analysis technique to help filter out underlying randomness or noise. Filters attempt to find the most likely signal that generated the series of observations.

For the purposes of this review, polynomial regression, piecewise polynomials, splines and then B-splines will be examined as a method for implementing a basis of splines. Finally the topic of wavelets will be discussed. With wavelet analysis, the original signal is decomposed into a series of coefficients, which carry both spectral and temporal information of the original signal. A discussion will follow to highlight the limitations and advantages of these methods.

Throughout the review, 1 minute epoch data from a single individual will be used to illustrate the various models. For the fitting of curves, a least squares approach will be used.

2.1 Polynomial Regression

Regression (Draper & Smith, 2014; Weisberg, 2005) is a way to describe the relationship between an dependent variable and one or more independent input variables. It answers questions about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor or a treatment on the value of the response (Weisberg, 2005). The simplest form of regression is linear, where a line is fitted to a set of data points. It describes an unchanging relationship between two phenomena. Mathematically, the line can be expressed as (Weisberg, 2005):

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.1}$$

where X is the independent input variable, Y is the dependent output variable and ϵ is the error term. The belief is that Y depends on X. β_0 is the intercept, the value of Y when X is zero. β_1 is the slope of the line, which characterises the relationship between the input and output. It is assumed that the error terms are independent and identically distributed with an expected value of zero and constant variance σ^2. For the remainder of this chapter, this error term is assumed to exist and will not be explicitly stated in equations. If there is more than one input variable which can be used to determine the output, then the linear expression (2.1) can be extended as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{2.2}$$

The relationship may not always be linear, and so in polynomial regression (Draper & Smith, 2014; Weisberg, 2005) the model is extended by including higher order terms. More formally polynomial regression can be defined as follows: A form of regression analysis in which the relationship between the independent variable X and the dependent variable Y is modelled as an n^{th} degree polynomial in x. A polynomial of degree D is a function formed by the linear combination of the powers of its argument up to D (Rawlings, Pantula, & Dickey, 2001):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_D X^D \tag{2.3}$$

This equation is for a single independent input variable, with various transforms applied.

These functions are defined globally, meaning that they apply across the full range of data. If the data has high variance, the function will be complex even

if some part of the data is constant or linear. To demonstrate how well polynomial regression can be used to represent the data, 1st, 3rd, 5th and 7th degree polynomials are fit to the data in Figure 2.1.

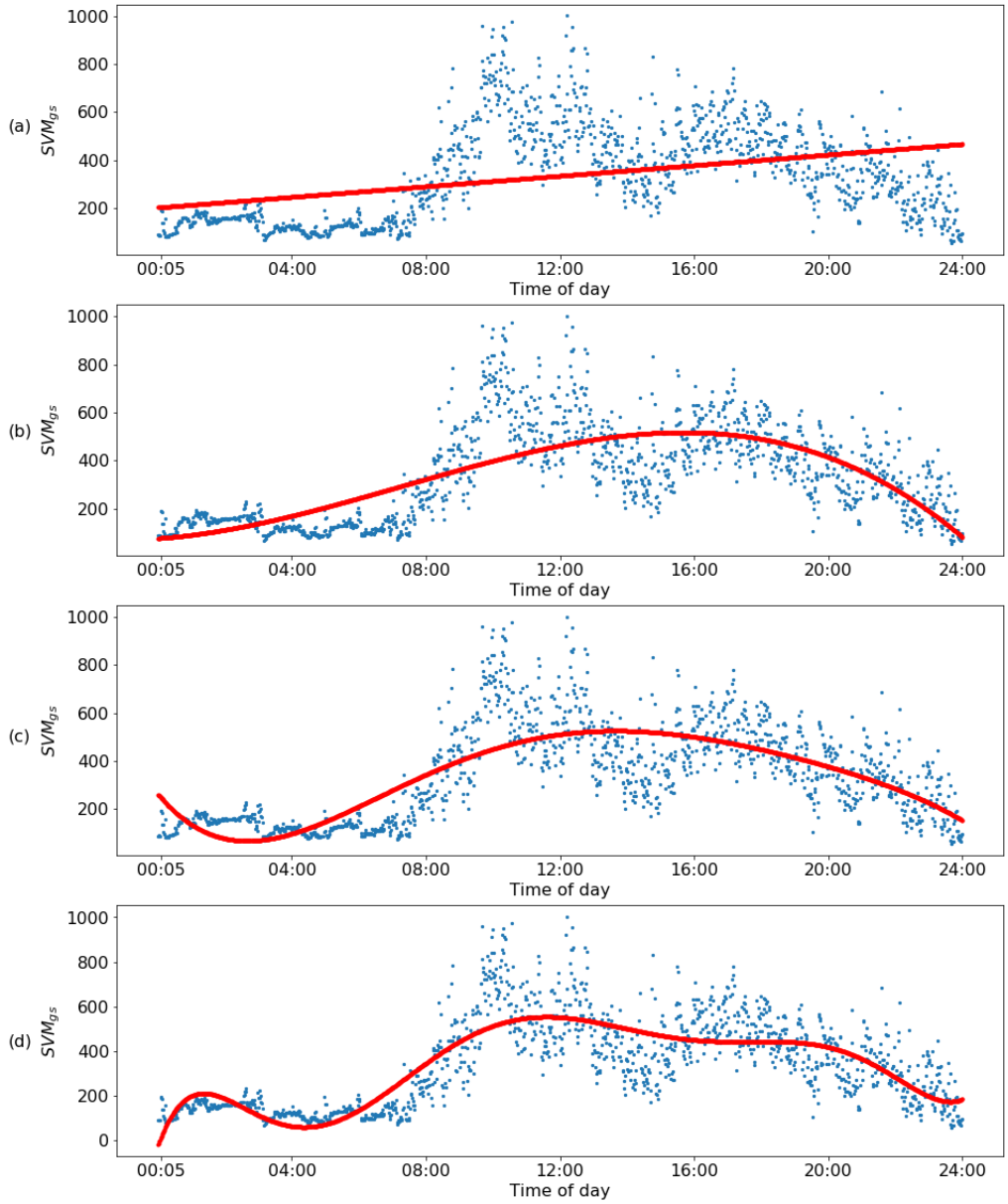


Figure 2.1: Polynomials with increasing degree: (a) 1st, (b) 3rd, (c) 5th, (d) 7th

As evident from Figure 2.1(a), the relationship is not linear. Increasing the order of the polynomials gets closer to describing the underlying nature of the data. Figure 2.1(b), while not fitting the data very well, still captures the intrinsic property that people are less active in the morning, increase activity during the day before decreasing again in the evening.

It is helpful to have a way to describe variability when discussing different models or smoothing techniques, so that these models can be compared when the complexity increases or when restrictions are placed (Friedman, Hastie, & Tibshirani, 2001). The number of *degrees of freedom* is the number of values in a calculation that can vary. In other words, it is the number of independent ways in which a dynamic system can move, without violating any constraint imposed upon it. So the linear model will have two degrees of freedom, as the intercept and slope can be varied to generate any line.

Polynomial regression has a number of benefits and limitations (Chambers, 2017). The benefits are its flexibility and interpretability, it is easy to understand and explain. However the coefficients themselves may not be easy to interpret. It also provides a good approximation. Typically the first tool used in data analysis to get a sense of any patterns in the data.

Limitations include under-fitting. The linear model in Figure 2.1(a) has under-fit the data. To overcome under-fitting, we need to increase the complexity of the model. This increases the number of features which can be difficult to handle. The polynomial with order 7 in Figure 2.1(d) still does not give a good approximation of the sample data. To get a better fit we need to increase the complexity of the model, namely by increasing the order of polynomial used, which could lead to over-fitting. Higher order polynomials should be avoided in regression as results based on high order polynomials are sensitive to the order of the polynomial (Gelman & Imbens, 2018). It is also inherently non-local. Changing the value of Y at one point can affect the fit of the polynomial for data points far away. To avoid the use of high degree polynomials on the whole dataset and to avoid their global nature, we can substitute in many small degree polynomials.

2.2 Piecewise Polynomials

Piecewise polynomials (Draper & Smith, 2014; Weisberg, 2005; Friedman et al., 2001) work by separating the data into different regions and then defining a different polynomial per region. Formally, a piecewise polynomial function, $f(X)$, is obtained by dividing the domain of X into contiguous intervals, and representing f by a separate polynomial in each interval (Friedman et al., 2001). The most simple of which is piecewise constant, or a polynomial with order zero.

The points of separation are known as knots. The knots cut the data into intervals or regions. These knots can be selected *a priori*, or we can allow the data to dictate. To illustrate, piecewise polynomials with ascending orders were fitted with two knots selected a priori, as seen shown in Figure 2.2. In this example the knots were placed uniformly.

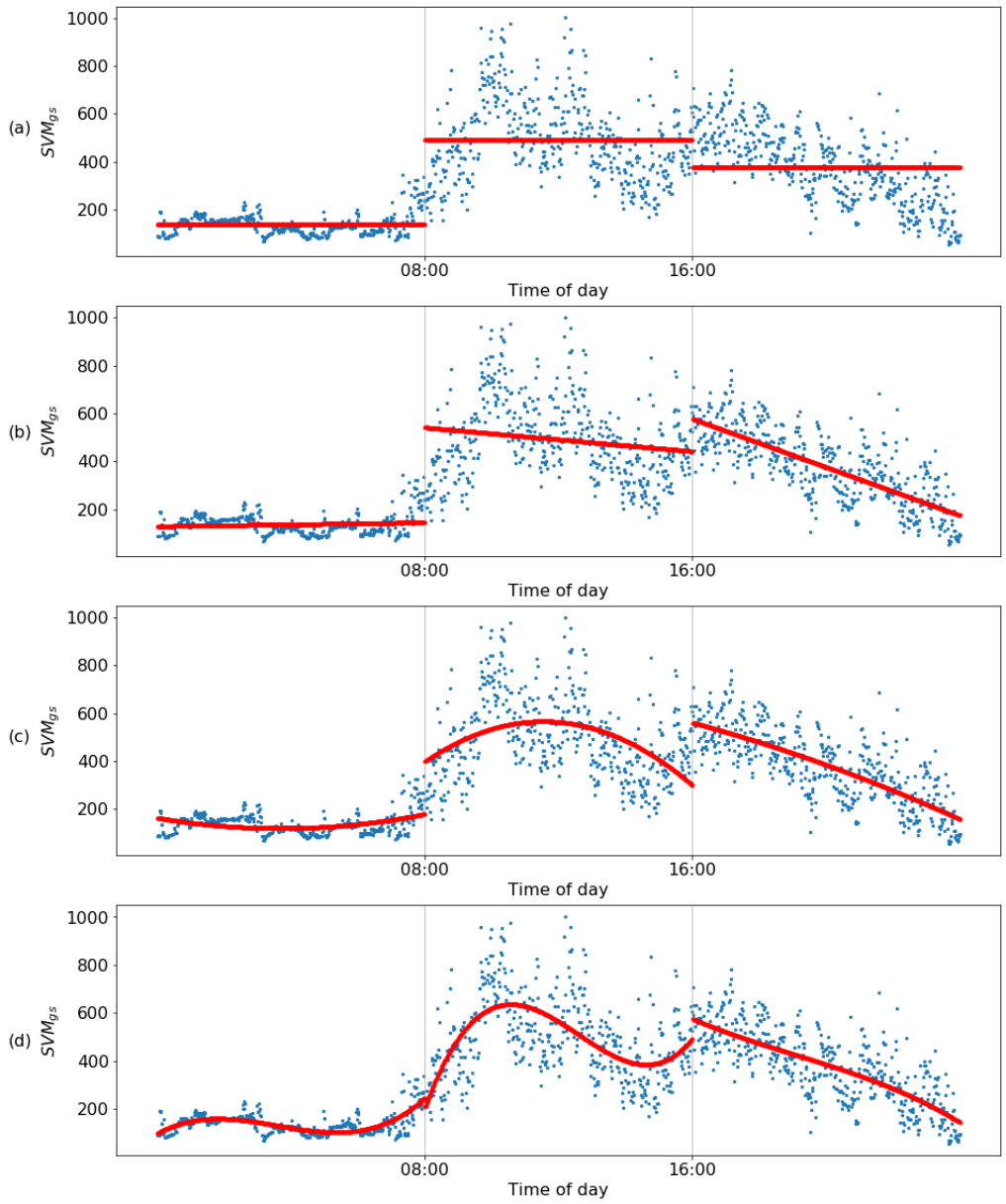


Figure 2.2: Piecewise polynomials with two knots and orders; (a) Zero, (b) One, (c) Two, (d) Three

The knots placed at 8a.m. and 4p.m. are represented by dashed vertical lines in Figure 2.2. Using more knots leads to a more flexible piecewise polynomial. As we use different functions in every interval, these functions will depend only on the distribution of data in that particular interval. The number of degrees of freedom (DoF) here, which do not have any constraints, can be calculated by:

DoF = (Number of regions) x (Number of parameters per region)

For example, a piecewise linear polynomial with two knots, will have six degrees of freedom.

The flaw with the implementation in Figure 2.2 is the discontinuities at the knots. This means that for a given input value, there are multiple outputs. A well-defined function associates one, and only one, output to any particular input so ideally every input should generate a unique output. The first constraint we can then place on this system is for it to be continuous at the knots. Doing this removes the ambiguity of the output value. Placing this constraint means that the function will have a unique output for every input, and generates graphs like those seen in Figure 2.3.

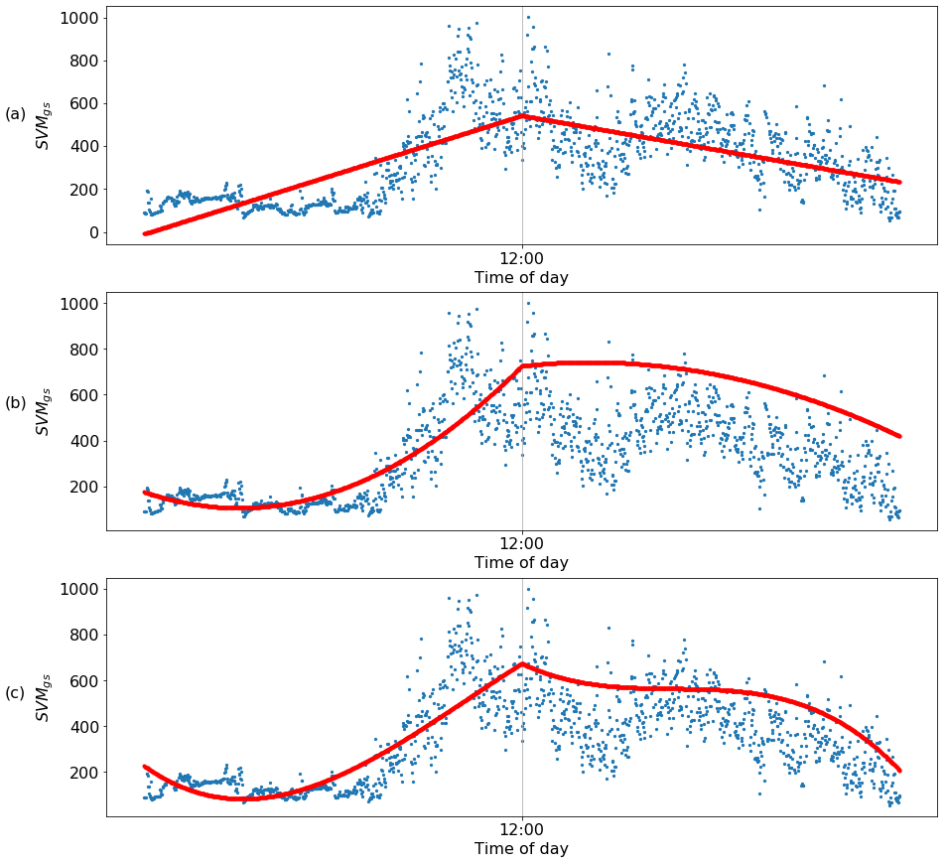


Figure 2.3: Piecewise continuous polynomials with one knot and increasing orders. (a) One, (b) Two, (c) Three

In this illustration one knot was selected a priori and polynomials of orders one, two and three were used. Continuing the example of a piecewise linear with two knots, we now have four degrees of freedom, as there is a constraint now on each of the knots. The calculation can then be extended to factor this is (Friedman et al., 2001):

DoF = ((Number of regions) x (Number of parameters per region)) - ((Number of knots) x (Number of constraints per knot))

Again these illustrations leave much to be desired, smoothness of the knots is still absent. To achieve this we need to add another constraint. Namely that the first derivative, i.e. the rate of change, of both polynomials either side of the knot must be the same. To illustrate the effect of this constraint, piecewise continuous linear polynomials for a single knot, with and without a continuous 1^{st} derivatives are shown in Figure 2.4 .

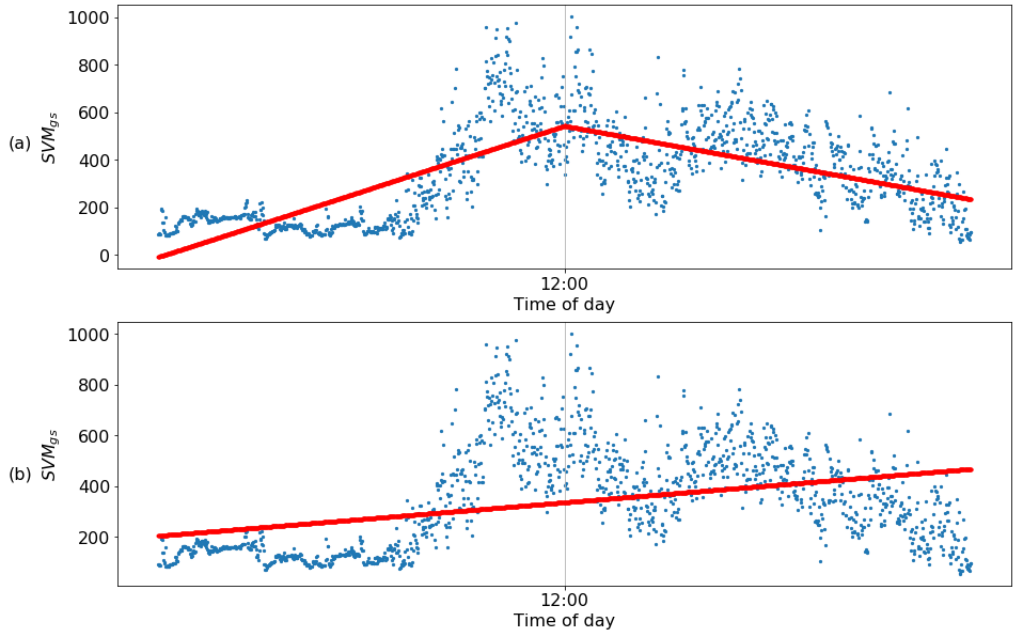


Figure 2.4: Piecewise continuous linear polynomials for one knot, (a) without continuous 1^{st} derivative, (b) with continuous 1^{st} derivative

Figure 2.4(b) is the same as performing linear regression, as the rate of change is constant throughout. A quick check of the parameter count confirms this:

DoF = ((2 regions) x (2 parameters per region)) - ((1 knot) x (2 constraints per knot)) = 2

This illustrates another point that to enjoy the benefits of constraining continuity at the knots, higher order polynomials are needed so that restrictions on the derivatives can be introduced. To achieve further smoothness another constraint can be imposed such that the 2^{nd} derivative, or the rate of change of the rate of change, is also continuous. This does not make any sense in the context of linear piecewise polynomials, since the 2^{nd} derivative does not exist, but for higher order polynomials the benefits can be seen.

Piecewise polynomials have addressed the limitation of polynomial regression being non-local but continuity then becomes an issue. This brings us to the topic of splines. The continuity in all of their lower order derivatives is what makes splines very smooth. It is claimed that cubic splines are the lowest-order spline for which the knot discontinuity is not visible to the human eye (Friedman et al., 2001).

2.3 Splines

A spline (Friedman et al., 2001; Wasserman, 2007) is a special piecewise polynomial. It consists of polynomial pieces on subintervals joined together with certain continuity conditions. An M^{th} order spline is a piecewise $M-1$ polynomial with $M-2$ continuous derivatives at the knots (Wasserman, 2007).

A linear spline is a continuous function formed by connecting linear segments. At its most simple, with first degree polynomials and the number of knots equal to the number of data points, a linear spline is the same as simple interpolation. Quadratic splines would have continuous 1^{st} derivative, cubic splines would have continuous 1^{st} and 2^{nd} derivatives (or twice continuously differentiable) and so on. There is seldom any reason to go beyond cubic splines unless smooth derivatives are of interest (Friedman et al., 2001). This property is often of interest when dealing with mathematical problems of convexity and convergence.

A cubic spline is a piecewise polynomial with a set of extra constraints. Namely continuity, continuity of the first derivative and continuity of the second derivative. Generally, a cubic spline with K knots will have a total of 4+K degrees of freedom (Friedman et al., 2001). For example, a cubic spline with 4 knots:

DoF= ((5 regions) x (4 parameters per region)) - ((4 knots) x (3 constraints per knot)) = 8

Knots are usually chosen in uniform space (De Boor, De Boor, Mathématicien, De Boor, & De Boor, 1978). One way to do this is to specify the desired degrees of freedom and then calculate where to place the knots at uniform quantiles of the data. Another option is to try out different numbers of knots and see which produces the best representation. To illustrate the effect of different knot sequences, cubic polynomials with an increasing number of uniformly distributed knots are fitted in Figure 2.5.

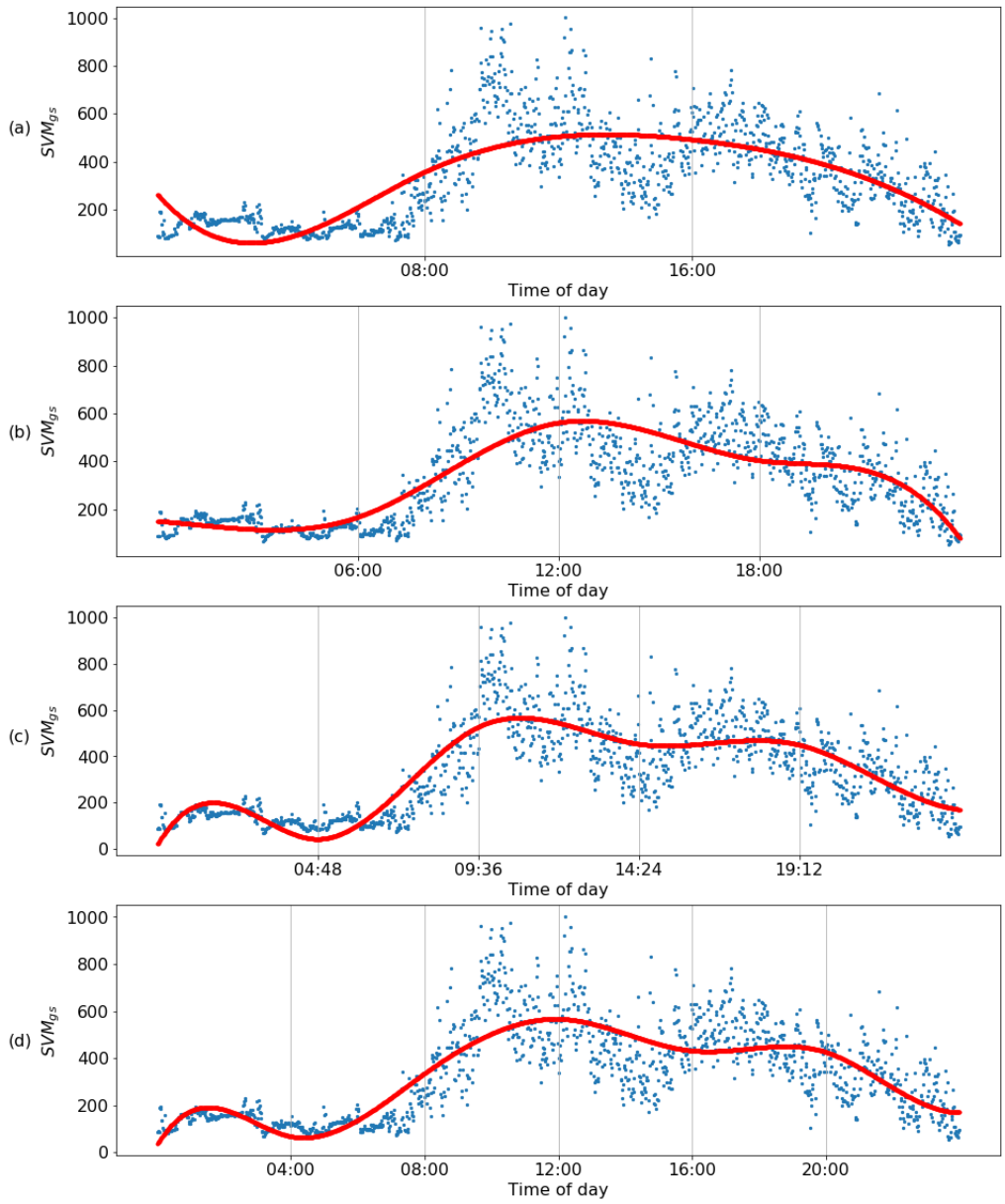


Figure 2.5: Cubic splines with uniformly distributed knots. (a) Two, (b) Three, (c) Four, (d) Five

The dashed vertical lines in Figure 2.5 again represent where the knots are placed. As the number of knots used increases, the curve appears to encapsulate more nuances of the data. Another potential solution for knot placement would be times in which there is high variability. At these times the polynomial coefficients can change rapidly. Hence, one option is to place more knots where the function might vary most rapidly, and to place fewer knots where it seems more stable. There appears to be little activity in the morning so fewer knots will be placed here, whereas in the middle of the day the activity seems to fluctuate so more

knots will be placed here, this is illustrated in Figure 2.6.

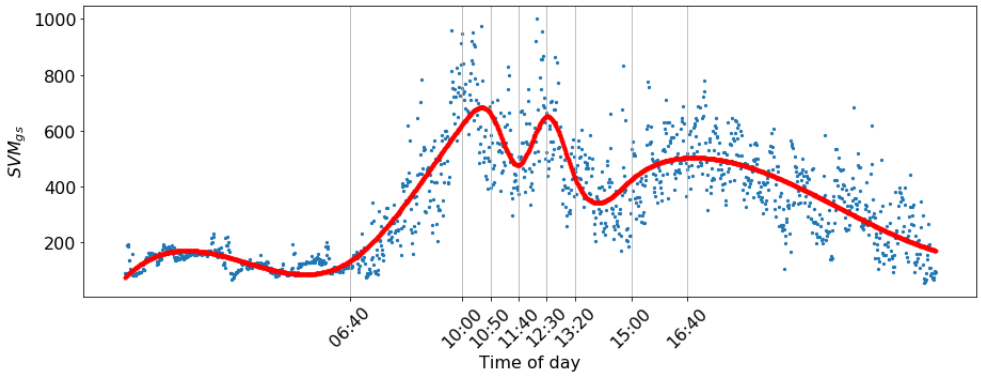


Figure 2.6: Cubic splines with knots at times of high variability

Eight knots were chosen manually, these are represented by dashed vertical lines in Figure 2.6 and the times at which they were placed are marked on the x-axis. The shape of a spline can be controlled by carefully choosing the number of knots and their exact locations in order to allow flexibility where the trend changes quickly. Figure 2.6 illustrates that placing knots where there is high variability generates a better representation of the data. It also allows us to avoid over-fitting where the trend changes little, as evident in the morning period in Figure 2.6.

Polynomial fit tends to be erratic near the boundaries. This issue is evident for polynomials, and becomes even more erratic when dealing with piecewise polynomials and splines. They behave erratically beyond their boundary knot points, and (typically) grow without bound outside of that range (Friedman et al., 2001). This instability makes extrapolation dangerous. To highlight this issue the cubic splines fitted in Figure 2.5, have been extrapolated beyond their boundaries, and are shown in Figure 2.7.

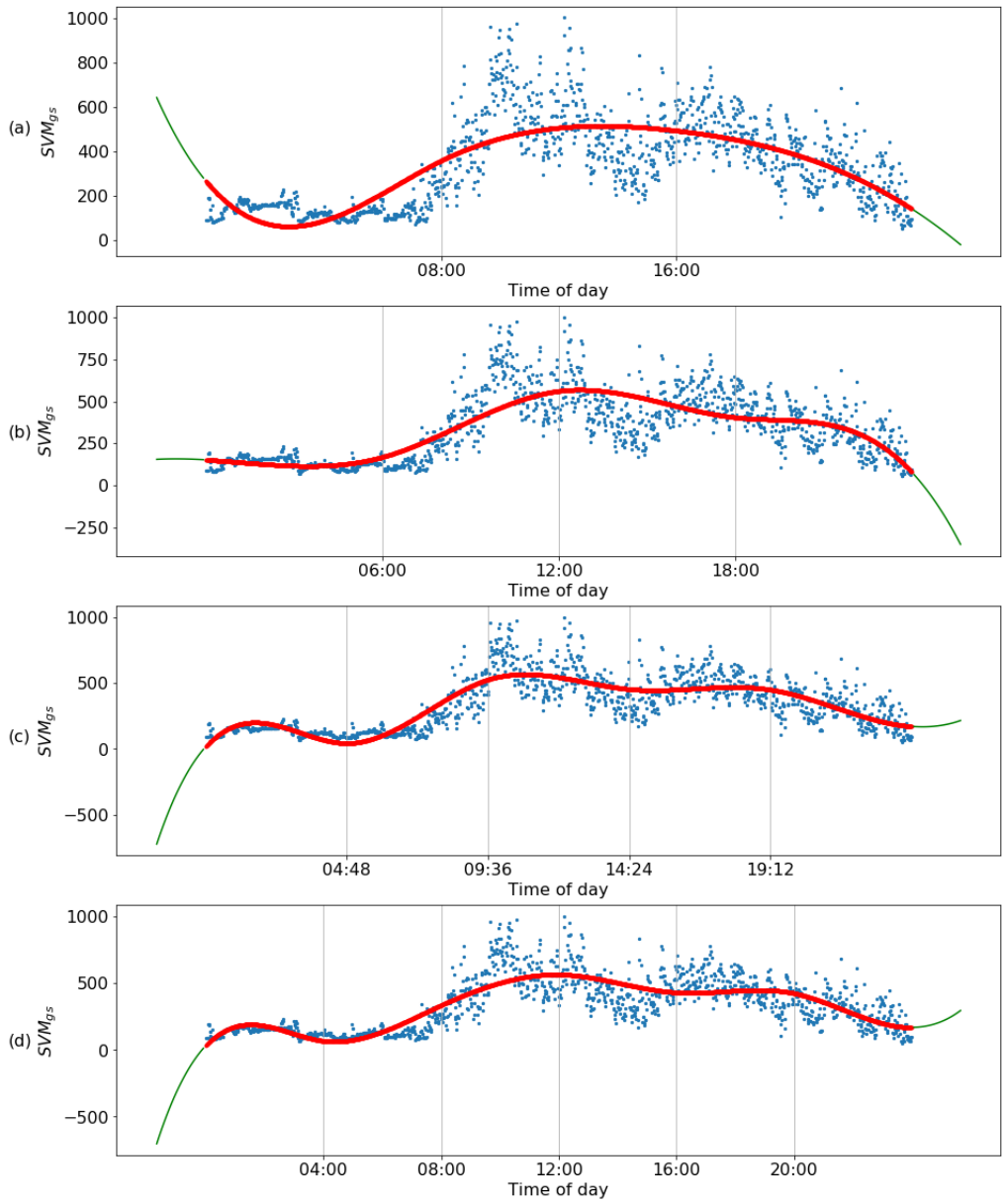


Figure 2.7: Cubic splines extrapolated beyond their boundaries with uniformly distributed knots. (a) Two, (b) Three, (c) Four, (d) Five

For example, after the right most data point in Figure 2.7(b) the function decreases below zero, which does not make sense in the context of activity. To address this issue a lower-degree polynomial can be used beyond the boundary knots. A spline that is linear beyond the boundary knots is called a natural spline (Wasserman, 2007). This adds additional constraints, namely that the function is linear beyond the boundary knots. This frees up four degrees of freedom (two from each boundary region), which can be spent more profitably by sprinkling more knots in the inner regions.

To capture non-linearity in regression models, we need to transform the input variables. A general *family* of transformations can be applied (Wasserman, 2013), that should be flexible enough to adapt to a wide variety of shapes, but not *too* flexible as to over-fit. This concept of a family of transformations is known as a basis of functions. Instead of fitting a linear function of powers of X, we fit the below model:

$$y_i = \beta_0 + \beta_1 h_1(x_i) + \beta_2 h_2(x_i) + ... + \beta_K h_K(x_i) \tag{2.4}$$

where h_K are the set of basis functions.

Representing the problem like this reduces it to finding the parameters $\beta_0, \beta_1, ..., \beta_K$. For linear regression, in this representation $h_1(x) = x$. Then in the general sense for polynomial regressions the choice of basis is to set $h_i(x) = x^i$ for i=1, 2, ..., D. By writing in this fashion the transformations are no longer limited to be of polynomial nature. Every spline can be represented by bases like these.

2.4 B-Spline

A basis for the set of natural splines that is particularly well suited for computation is the B-spline basis (Wasserman, 2007). B-spline curves are composed from many polynomial pieces, with the domain again subdivided by knots. Each B-spline basis function is non-zero on a few adjacent subintervals. To begin with, functions are defined piecewise constant:

$$B_{i,1}(x) = \begin{cases} 1 & \tau_i \leq x < \tau_{i+1} \\ 0 & otherwise \end{cases}$$

where $B_{i,m}(x)$ denotes the i^{th} B-spline basis function of order m for the knot sequence τ (Friedman et al., 2001).

For example, take a knot sequence of (00:00, 08:00, 16:00, 24:00) with the boundaries at [00:00, 24:00]. The basis functions are defined as:

$$B_{1,1}(x) = \begin{cases} 1 & 00:00 \leq x < 08:00 \\ 0 & otherwise \end{cases}$$

$$B_{2,1}(x) = \begin{cases} 1 & 08:00 \leq x < 16:00 \\ 0 & otherwise \end{cases}$$

$$B_{3,1}(x) = \begin{cases} 1 & 16:00 \leq x < 24:00 \\ 0 & otherwise \end{cases}$$

These functions are displayed in Figure 2.8.

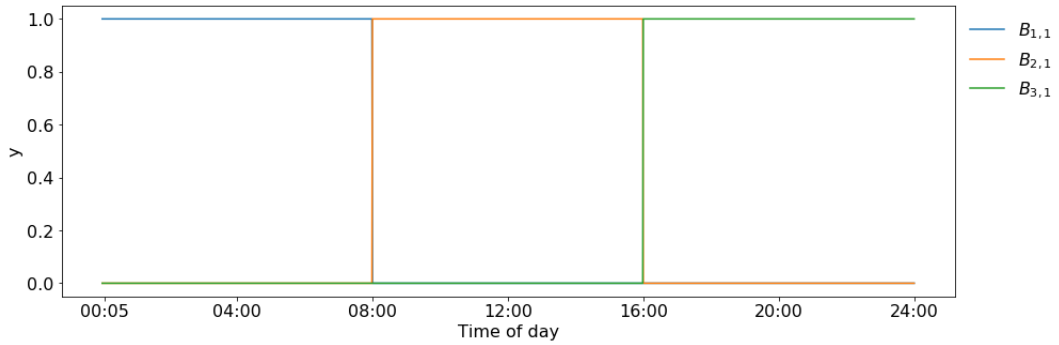


Figure 2.8: Basis functions for a B-spline with order 1

These are also known as Haar basis functions (Friedman et al., 2001). The B-splines should form a partition of unity, i.e.,

$$\sum_i B_{i,1}(x) = 1, \forall x \tag{2.5}$$

This constraint is necessary so that the basis functions are able to span the entire space (Ohtake, Belyaev, Alexa, Turk, & Seidel, 2003). The four knots (two boundary and two interior) create three intervals with a function defined for each.

Higher order splines are then defined iteratively by:

$$B_{i,m}(x) = \frac{(x - \tau_i)}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{(\tau_{i+m} - x)}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x) \tag{2.6}$$

where the $B_{i,m}$ are called the i^{th} B-Spline basis functions of order m, and the recurrence relation is called the De Boor recurrence relation (De Boor et al., 1978). For m=2, Equation (2.6) becomes:

$$B_{1,2}(x) = \frac{(x - \tau_1)}{\tau_2 - \tau_1} B_{1,1}(x) + \frac{(\tau_3 - x)}{\tau_3 - \tau_2} B_{2,1}(x) \tag{2.7}$$

This equation is now defined over two of the subintervals described by the knots. Figure 2.9(a) shows these basis functions. These functions, however, no longer form a partition of unity. To fix this, the knot sequence is augmented with additional knots at the boundaries.

The new sequence becomes (00:00, 00:00, 08:00, 16:00, 24:00, 24:00) and Equation 2.6) can be used to generate the basis functions shown in Figure 2.9(b).

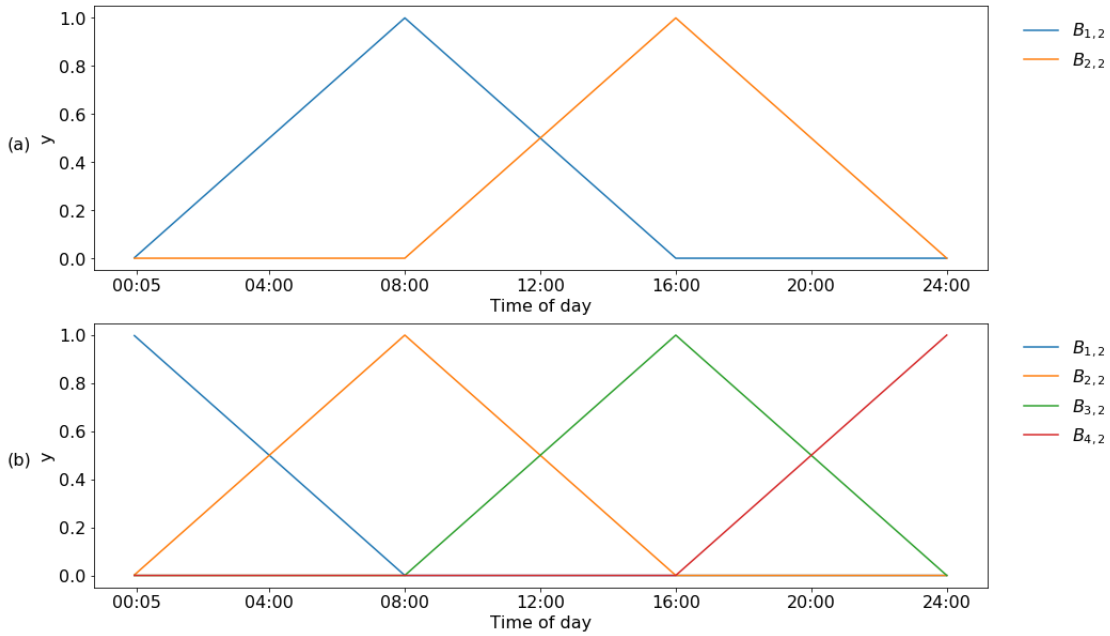


Figure 2.9: Basis functions for a B-spline with order 2. (a) without augmented knots, (b) with augmented knots

Extending this logic further with more augmentations and iterations we can generate the basis functions for quadratic and cubic B-splines, as shown in Figure 2.10.

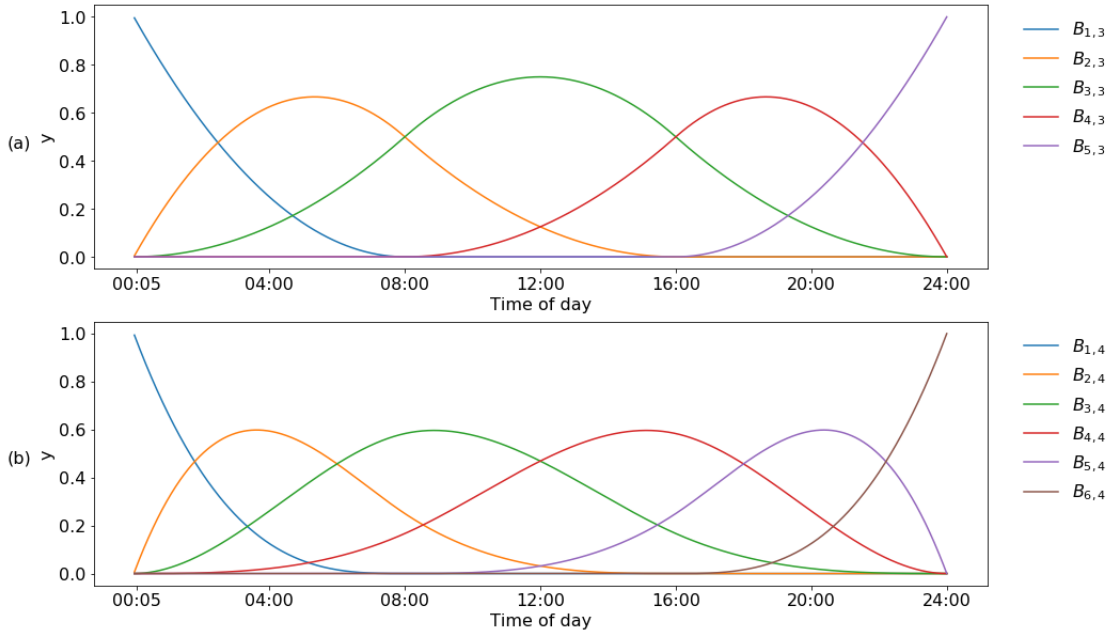


Figure 2.10: Basis functions for a B-spline with orders: (a) 3, (b) 4

With the basis functions now defined, a B-spline can then be defined as a linear

combination of these (Friedman et al., 2001). A B-spline of degree n (of spline order $m = n + 1$) is a parametric curve composed of a linear combination of basis B-splines $B_{i,n}(x)$ of degree n given by:

$$B(x) = \sum_{i=0}^{N+n} \beta_i B_{i,n}(x), \quad x \in [\tau_0, \tau_{N+1}] \tag{2.8}$$

where β_i are the coefficients of the linear combination and N is the number of interior knots.

An alternative approach to consider would be wavelets.

2.5 Wavelets

Wavelets (Strang & Nguyen, 1996; Walnut, 2013) are basis functions that can be used to represent other functions. At their most simple, wavelets are like mini waves. They wave above and below the x-axis and integrate to zero. Unlike sine or cosine which continue forever, wavelets are a short burst of waves that quickly die away. Another way to describe this is that the wavelet has compact support, meaning that the signal does not last forever, or that the function is non-zero on a limited portion of its domain. Sines and cosines are by their definition not local and therefore do a poor job at approximating sharp spikes (Graps, 1995). Wavelets come in many different forms and are used as the basis functions in a wavelet transform, examples of wavelets can be seen in Figure 2.11.

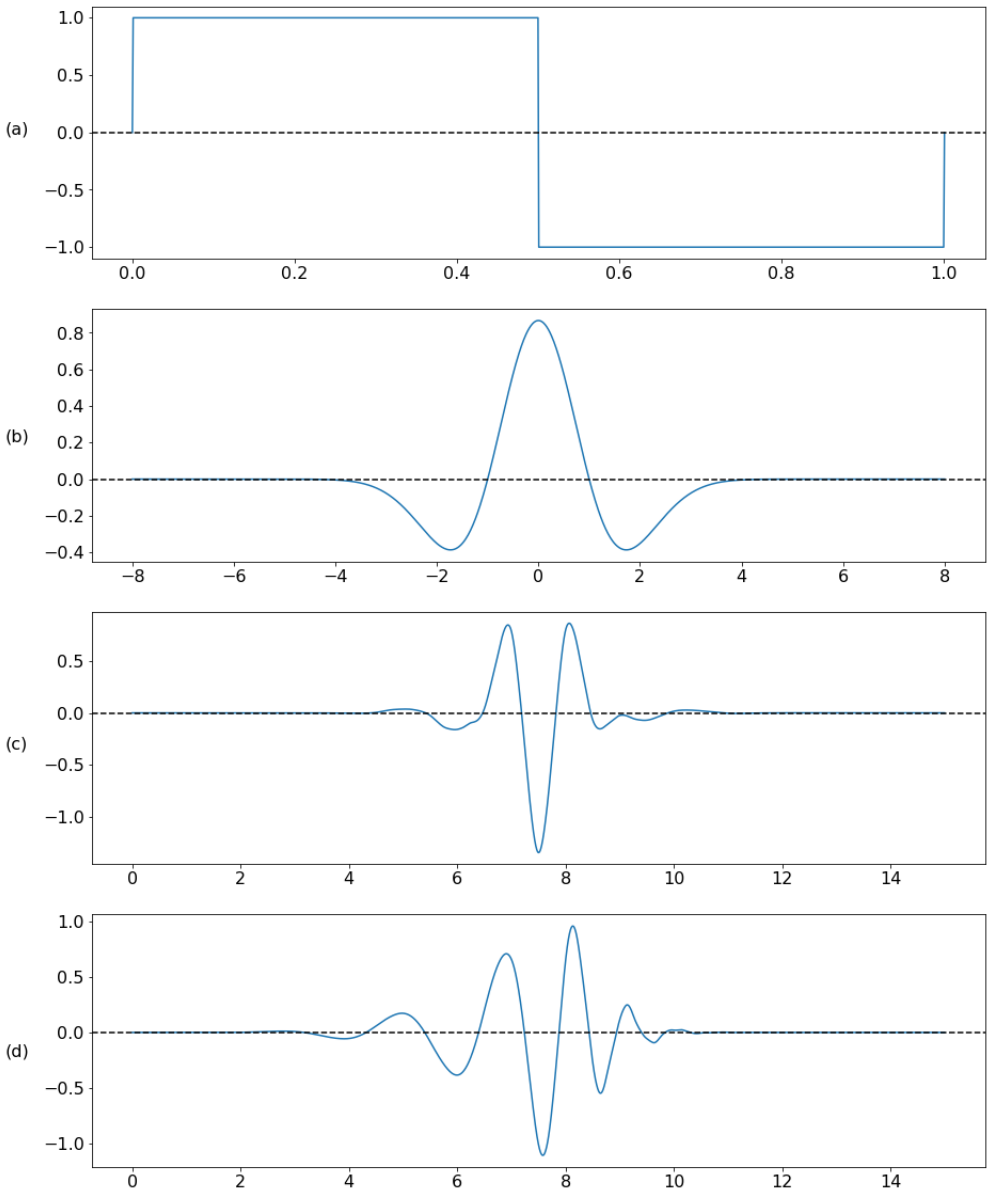


Figure 2.11: Wavelet examples (a) Haar, (b) Mexican hat, (c) Symmlet, (d) Daubechies

Functions can be approximated by scaling and shifting these wavelets. The Haar mother wavelet is defined by (Walnut, 2013):

$$\varphi(x) = \begin{cases} 1 & 0 \leq x \leq \frac{1}{2} \\ -1 & \frac{1}{2} \leq x \leq 1 \end{cases}$$

Once φ is fixed, translations and dilations can be formed as follows. For $k \in \{0, 1, \ldots, 2^j\}$, define:

$$\varphi_{j,k}(x) = 2^{\frac{j}{2}} \varphi(2^j x - k) \tag{2.9}$$

The function $\varphi_{j,k}$ has the same shape φ but it has been rescaled by a factor of $2^{\frac{j}{2}}$ and shifted by a factor of k (Wasserman, 2013). Therefore all the wavelets are generated from a single basic wavelet, the mother wavelet. To illustrate these shifts (translations) and scales (dilations), the first 3 scales are shown in Figure 2.12. At each scale the wavelets are packed in side by side to completely fill the time axis, i.e. all translations at each scale are shown.

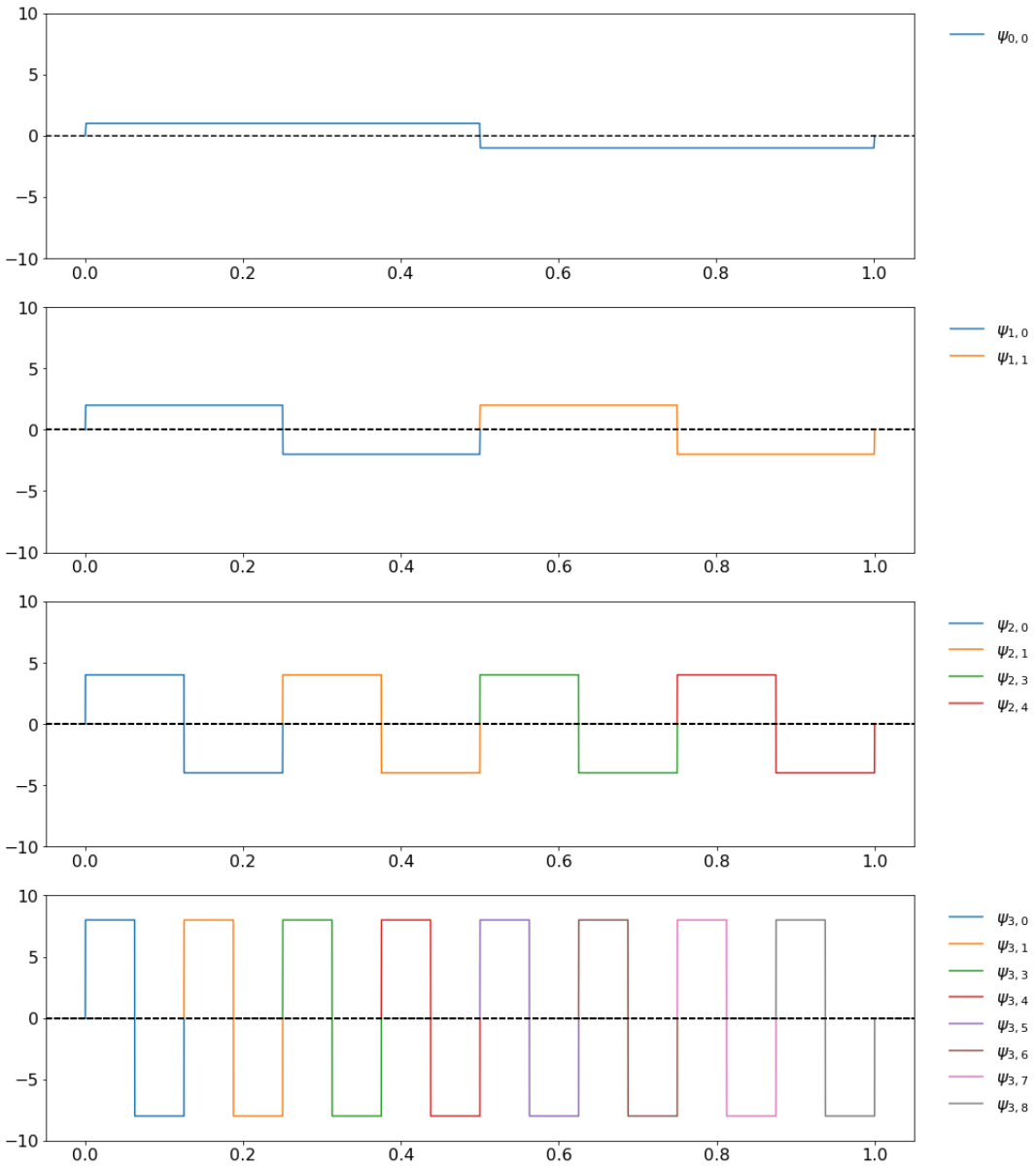


Figure 2.12: Select translations and dilations of the Haar wavelet family

Wavelet transforms typically use a complete orthonormal (they are both orthogonal and normalized) basis to represent functions, but then shrink the coefficients toward a sparse representation (Friedman et al., 2001). Two

functions are said to be orthogonal if their inner product is zero. Put simply this means that the area above and below the x-axis are the same. For example, $\varphi_{1,0}$ and $\varphi_{1,1}$ in Figure 2.13 are orthogonal. Normalized typically means that it is of unit length. The constant that makes the orthogonal basis orthonormal is $2^{\frac{j}{2}}$ in Equation (2.9).

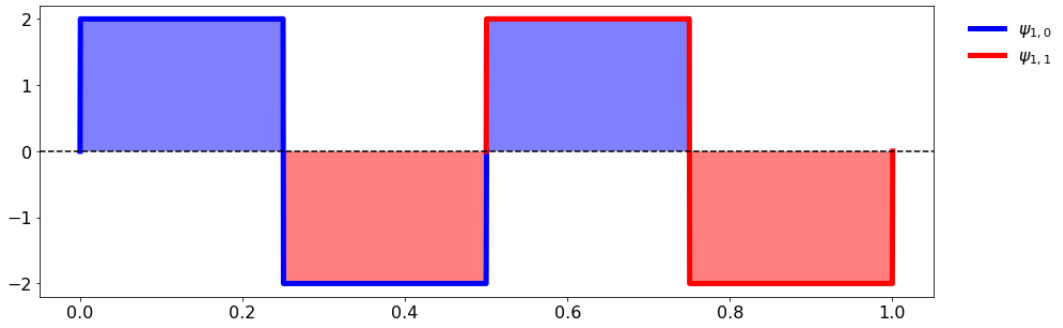


Figure 2.13: Orthogonality example

The large scales at the top of Figure 2.12 can be used to define the bigger picture, whereas the small scales show the details. Wavelet algorithms process data at different scales. If we look at data with a large window, we would notice gross features. Similarly, if we look at data with a small window we would notice small features. The result in wavelet analysis is to see both the forest (gross) and the trees (small) (Graps, 1995).

The set of all $\varphi_{j,k}$, however, does not form a complete basis to define any function in this space. Every time we scale a wavelet by a factor of two, the bandwidth is halved, which means that an infinite number of wavelets would be required to span the whole space. To span our data domain at different resolutions, the mother wavelet is used in a scaling equation. This scaling function, $\phi(x)$, is also known as the father wavelet (Friedman et al., 2001). The Haar father wavelet is defined by (Walnut, 2013):

$$\phi(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & otherwise \end{cases}$$

This father wavelet too can be scaled and translated, and so combined with the dilations of the mother wavelet to form the orthonormal basis for a Haar wavelet transform. This father wavelet is shown in Figure 2.14.

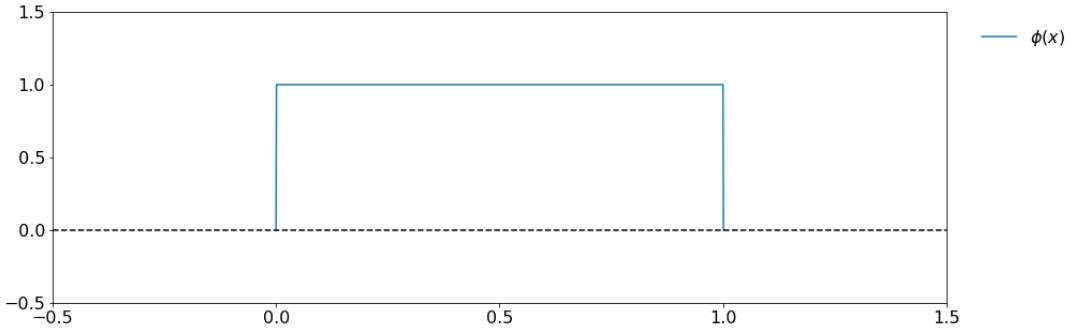


Figure 2.14: Haar father wavelet

Translations and dilations can also be taken of this father wavelet, similarly to how it is performed for the mother:

$$\phi_{j,k}(x) = 2^{\frac{j}{2}}\phi(2^j x - k) \tag{2.10}$$

Wavelet methods work by representing a curve as a sum of functions, which are all scaled and time shifted versions of the mother and father wavelets. While scaling of the mother wavelet enables frequency resolution, the shifting provides the time information. Using a combination of the mother and father wavelets, a curve can be decomposed into distinct components. One such method for doing this is the Discrete Wavelet Transform (DWT).

The DWT can be interpreted as a filter bank, where the curve is decomposed into several components each representing a single frequency sub-band of the original curve. More details of these methods can be found in Strang and Nguyen (1996). DWT works by extracting multi-scale information from the sequence. Given two adjacent observations in the sequence, an approximation of these can be calculated by taking the average, and the degree of difference can be calculated by simple subtraction. This degree of difference can be thought of as the detail. By doing these calculations pairwise along the entire sequence, we can extract the approximations and details in the sequence at different scales and locations. Multi-scale means that we can then perform these decompositions again to obtain coarser information, approximations of the approximations, and the degree of difference in these approximations. Figure 2.15 shows this decomposition tree. This process is repeated to the desired composition level.

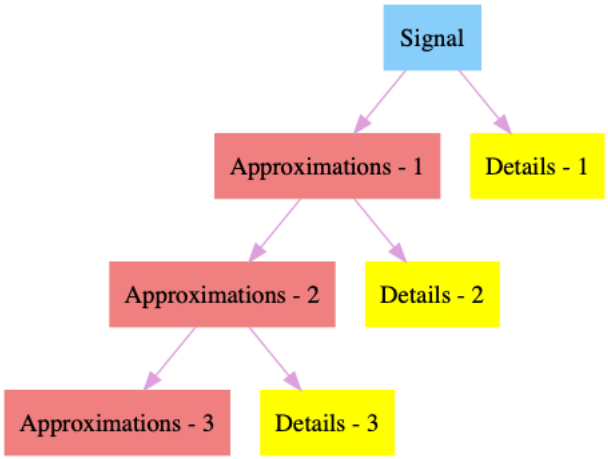


Figure 2.15: Wavelet decomposition tree

Taking the average, or fitting a constant function between points, is how the Haar wavelet transform calculates these approximations. Figure 2.16 shows the Haar wavelet transform fitted to a sequence of data with increasingly coarser scales and the details removed. Figure 2.16(a) calculates the average of two adjacent observations, then four, eight and 16. This can be viewed as the bigger picture (forest) mentioned earlier, and with each iteration we lose more and more detail.

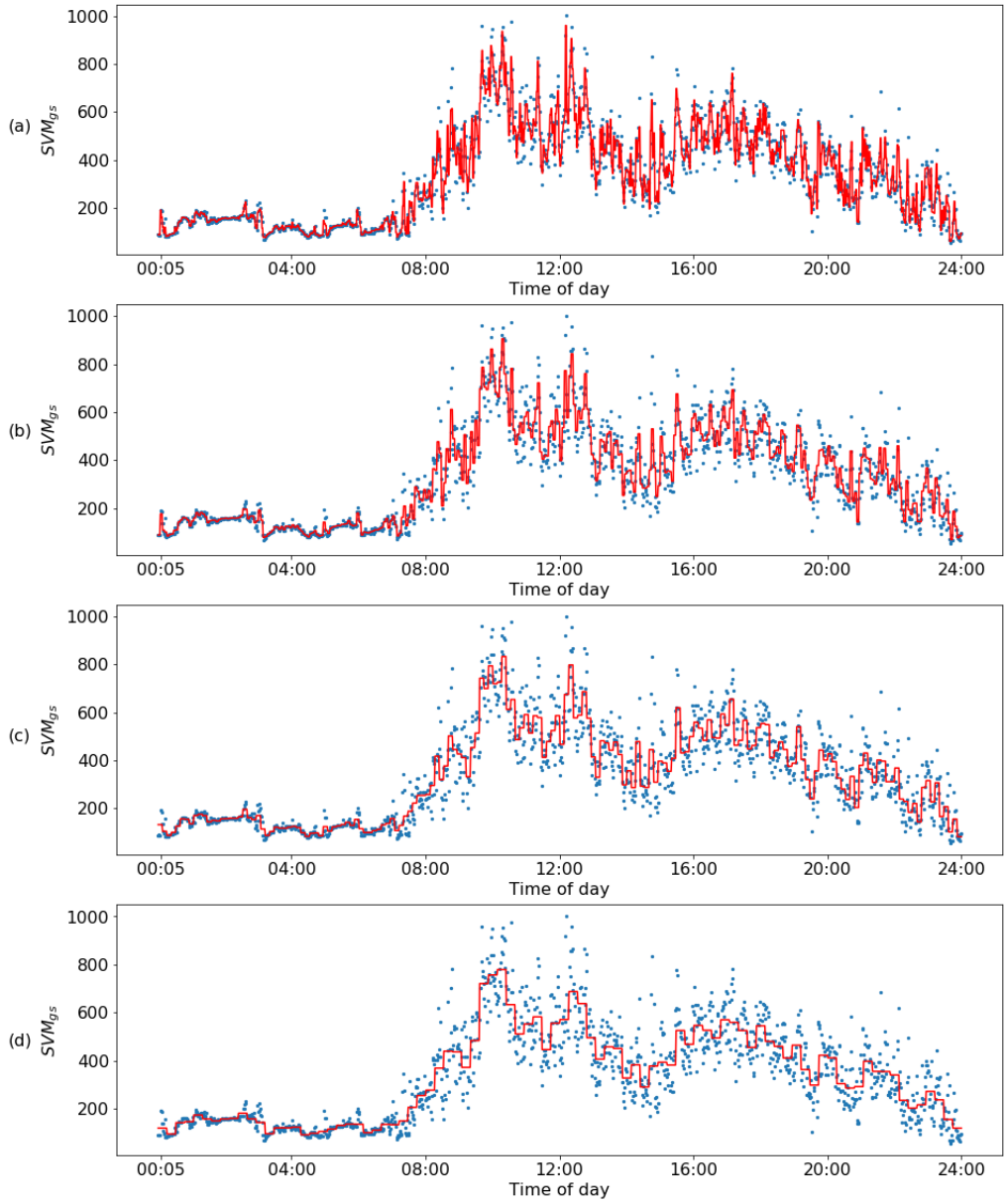


Figure 2.16: Scale approximations. (a) 1st, (b) 2nd, (c) 3rd, (d) 4th

These can be viewed as linear combinations of dilations of the father wavelet in Figure 2.14. The mother wavelets are responsible for the details at each level. The first decomposition of the sequence and the details are shown in Figure 2.17 . Figure 2.17(c) thresholds the differences with some arbitrary number to highlight the biggest differences. These are the small features, the trees in the forest.

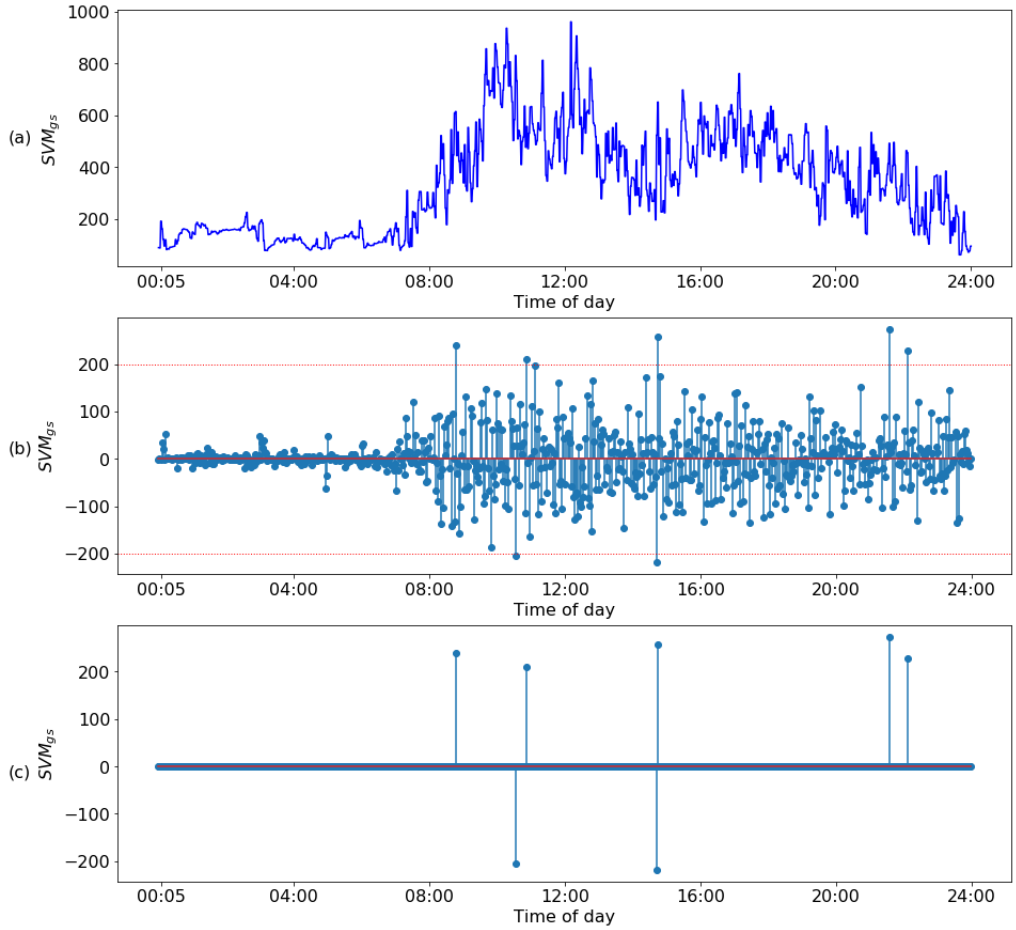


Figure 2.17: Haar with details (a) 1st scale approximation, (b) Raw details, (c) Thresholded details

In general, a wavelet series approximation to a continuous function $f(t)$ is given by (Morris et al., 2006):

$$f(t) = \sum_{k} c_{J,k}\phi_{J,k}(t) + \sum_{j=1}^{J}\sum_{k} d_{j,k}\varphi_{j,k}(t) \tag{2.11}$$

where J is the number of scales, and k ranges from 1 to K_j, the number of coefficients at scale j. The functions $\phi_{J,k}(t)$ and $\varphi_{j,k}(t)$ are wavelet basis functions that provide a location-scale decomposition of the observed function. They are dilations and translations of the father and mother wavelets respectively. The coefficients $c_{J,k}$ are the smooth coefficents, or the approximations. The coefficients $d_{J,k}, ..., d_{1,k}$ are the detail coefficients and represent deviations of the function at scale j, where smaller j correspond to finer scales.

Wavelets can be used to perform non-parametric regression using the following procedure:

1. Noisy data is projected into the wavelet domain using the DWT, which yields the wavelet coefficients $d_{J,k}, ..., d_{1,k}$.

2. These coefficients are then thresholded, by setting to zero any coefficients lower in magnitude than a specified threshold. This provides estimates of the true wavelet coefficients, which are the coefficients for the true function if there was no noise.

3. These estimates are then projected back to the data domain using the Inverse DWT (IDWT), which provides a de-noised non-parametric estimate of the true function, that retains dominant local features. This property makes the procedure useful for modelling functions with many local features like peaks.

A disadvantage of the Haar wavelet is that it is not smooth, but its simplicity means it is a good example for introducing wavelets. If the curve under consideration was discontinuous in nature, then Haar wavelets might be useful.

One way we can differentiate between wavelets is by the number of vanishing moments. The k^{th} moment of a function f is defined as the integral of the function multiplied by its variable to the power of k:

$$m_k = \int_{-\infty}^{\infty} f(x) x^k dx \tag{2.12}$$

The k^{th} moment vanishes if this integral is zero. A wavelet with a higher number of vanishing moments is more complex and is better able to accurately represent a complex signal. The price that is typically paid for more vanishing moments is a wider support. As the number of vanishing moments increases, polynomials up to that order will not be identified by the wavelet. For example, the Haar wavelet has one vanishing moment.

The more vanishing moments in the wavelet, the higher the regularity. Wavelets with low regularity create jagged representations of the data, see Figure 2.16. Using wavelets with higher regularity produces smoother representations of the function.

Daubechies (1992) constructed compactly supported orthogonal wavelets with a pre-assigned degree of regularity/smoothness. These wavelets are usually defined by their number of vanishing moments (m). The Haar wavelet is a special case of the Daubechies, with $m = 1$. Figure 2.18 illustrates Daubechies mother and father wavelets with increasing numbers of vanishing moments.

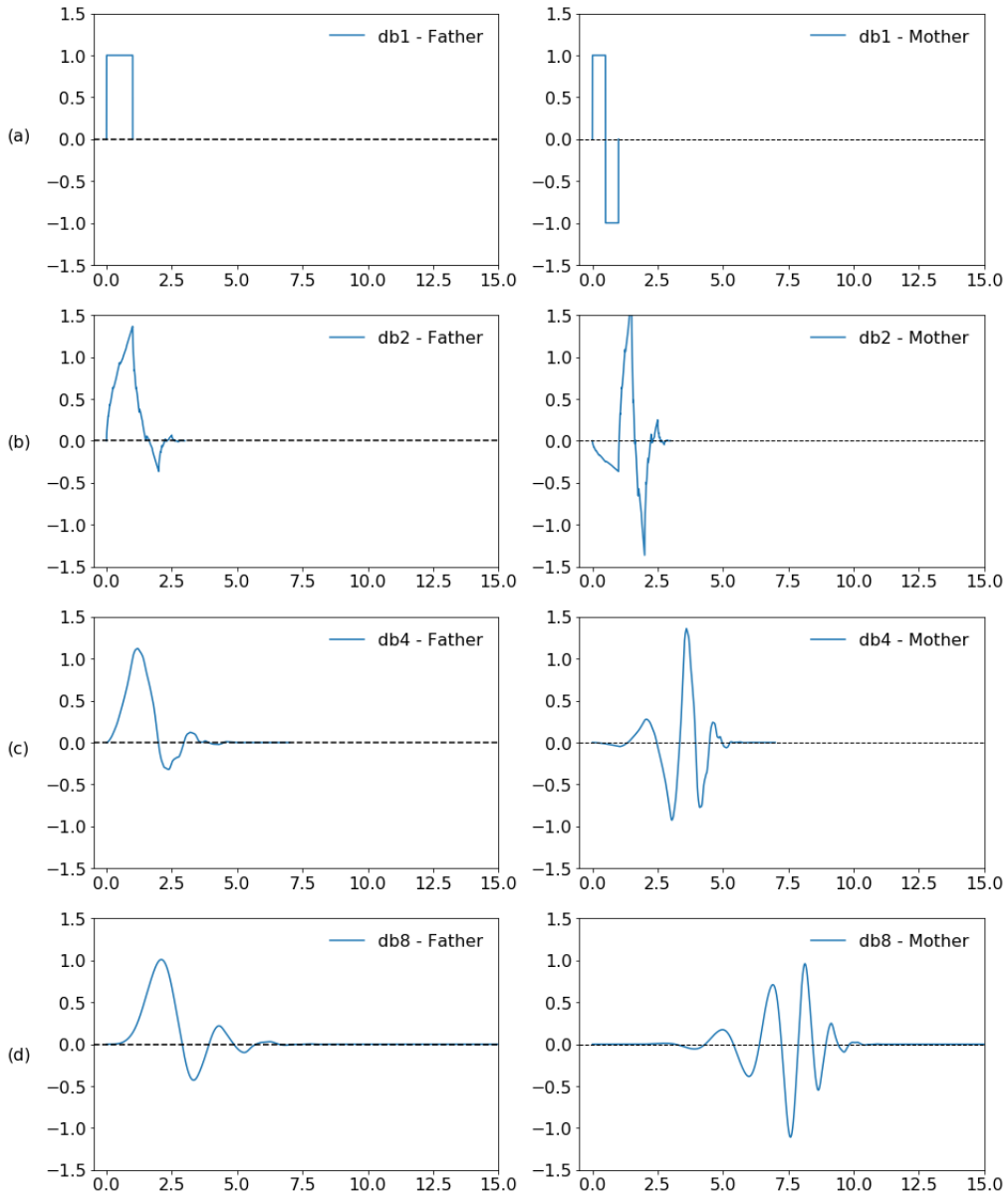


Figure 2.18: Daubechies mother and father. (a) m=1, (b) m=2, (c) m=4, (d) m=8

Choosing a Daubechies wavelet with four vanishing moments and performing wavelet decomposition yields the approximations shown in Figure 2.19. These wavelet transforms are carried out in the same way as the Haar, by computing approximations and differences. They only differ in how these scaling signals and wavelets are defined. Daubechies have slightly longer supports, meaning they produce these approximations and differences using more values from the signal. When compared to Figure 2.16, these approximations are a lot smoother.

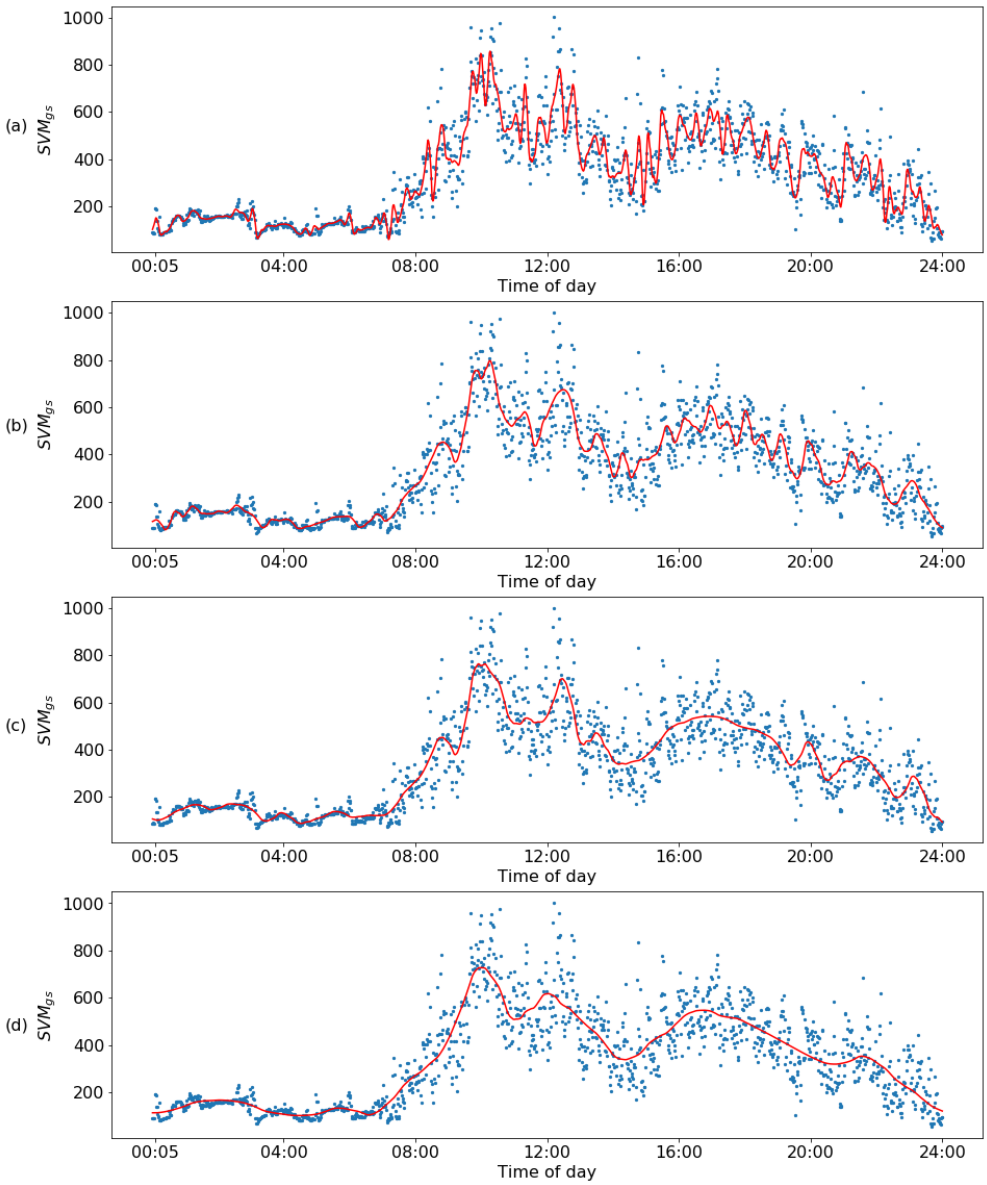


Figure 2.19: Daubechies-4 DWT with increasing scale approximations. (a) 3^{rd}, (b) 4^{th}, (c) 5^{th}, (d) 6^{th}

Unlike the DWT, the continuous wavelet transform (CWT) can operate at every scale. The CWT involves convolving a signal with an infinite number of functions, which are generated from translations and scaling of a specified mother wavelet function. CWT does not have a father scaling function. The resulting transform has parameters that vary continuously, meaning the inverse transform requires an infinite number of coefficients. However, since the curve under consideration is given in a discrete setting, we do not necessarily need smoothly varying parameters to reconstruct the signal from the coefficients. A DWT is sufficient for the curves, hence the CWT will not be considered.

2.6 Summary

The individual activity profile used for illustrative purposes in this review could not be represented by a simple parametric structure. Regression splines often give better results than polynomial regression (Friedman et al., 2001). This is because, unlike polynomials, which must use a high degree polynomial to produce flexible fits, splines introduce flexibility by increasing the number of knots but keep the degree fixed. Generally this approach produces more stable estimates. The problem is not that more knots are better than fewer knots, it is that the variables under consideration do not behave like that. Splines are not well suited for modelling functional data with many local features like peaks (Morris & Carroll, 2006).

Wavelet bases are chosen to represent data displaying discontinuities and/or rapid changes in behaviour (Ruppert, Wand, & Carroll, 2009). A wavelet transform is used to deconstruct the signal into a number of wavelets being added together. Similar to how a smooth function can be represented by a few spline basis functions, a mostly flat function can, with a few isolated bumps, be represented with a few bumpy basis functions. Wavelets are popular as they are able to represent smooth and/or locally bumpy functions in an efficient way (Friedman et al., 2001). Wavelets are well-suited for approximating data with sharp discontinuities.

From an initial exploration of the data, it is evident that these sharp discontinuities are present and so wavelets would be a good choice of smoothing method. Additionally, wavelets do not have the problem of knot selection evident in splines. To compare these methods presented in this chapter objectively, the predicted values from the smoothing technique can be compared to the actual observations.

These differences are prediction errors, and can be measured by the vertical differences between the actual values and the fitted line. The mean square error (MSE), (Friedman et al., 2001) is an average of the spread of the data around the fitted line, and reflects how big the typical prediction error is. The MSE was calculated for selected methods applied to every individual in the cohort and the minimum, mean and maximum of these values is presented in Table 2.1.

Polynomial Regression	Min MSE	Mean MSE	Max MSE
Order:1	2627	56519	231715
Order:3	2503	34287	146005
Order:5	2425	29426	132488
Order:7	2299	27166	119842
Cubic Splines			
2 Uniform knots	2438	29165	127742
3 Uniform knots	2266	28268	132564
4 Uniform knots	2296	26698	121957
5 Uniform knots	2238	25926	116510
11 Uniform knots	2051	21354	93165
23 Uniform knots	1923	16808	66879
Wavelets			
Level: 4			
haar	1261	11110	39394
db4	1200	10070	35760
db8	1406	10044	37475
Level: 5			
haar	1714	14750	56515
db4	1764	13394	53252
db8	1752	13303	49734
Level: 6			
haar	1891	19134	85397
db4	1922	17431	67107
db8	1927	17309	71223

Table 2.1: MSE for smoothing techniques

As suspected when applied to an individual, linear regression has the worst performance for the cohort. Using this metric, wavelets has outperformed the other methods, except for the cubic spline with 23 uniform knots (i.e. placed every hour) which has similar performance.

The main challenge in using wavelet transforms is to select the optimum mother wavelet for the given tasks, as different mother wavelets applied to the same signal may produce different results (Ngui, Leong, Hee, & Abdelrhman, 2013). Mother wavelets are characterised by properties including orthogonality, compact support, symmetry and the number of vanishing moments. These properties are considered when selecting the optimum mother wavelet to use; however, more than one mother wavelet with the same properties often exists. In this case, the similarity between the signal and the wavelet could be considered.

Figure 2.20 shows a comparison between Daubechies wavelets with four and

eight vanishing moments (see Figure 2.18 for the mother wavelets).

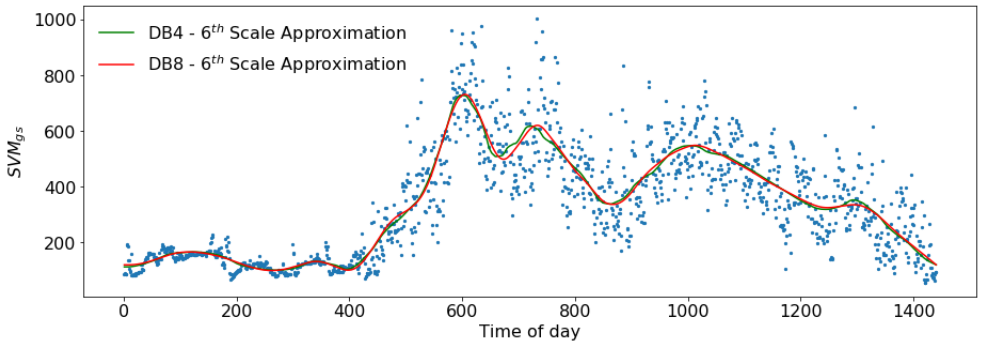


Figure 2.20: 6^{th} scale Daubechies - 4 and 8 WT

Visual inspection suggests that there is not much difference between the Daubechies-4 and Daubechies-8 mother wavelets to use. Daubechies-4 has fewer vanishing moments, and therefore more compact support. Therefore, this wavelet will be chosen for use in the next chapter which will utilise cluster analysis to identify subgroups within the cohort.

Chapter 3 - Cluster Analysis

Cluster analysis (Jain & Dubes, 1988; Hair et al., 2006) groups objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. The attempt is to maximize the homogeneity of objects within clusters while also maximizing the heterogeneity between clusters (Hair et al., 2006; Jain & Dubes, 1988).

Other studies (Lee, Yu, McDowell, Leung, & Lam, 2013; Staudenmayer, Pober, Crouter, Bassett, & Freedson, 2009) were explored in Chapter 1, that have clustered people based on physical activity. This provided the motivation for the belief that subgroups do exist within the cohort.

The next section, Methods, will discuss the components of clustering. In the presence of measurement errors, applying clustering methods directly to raw data does not take advantage of the functional structure. Thus, it is advantageous to partition the data while keeping the functional structure. This can be done by fitting curves for every individual, such as those discussed in Chapter 2, before proceeding with the cluster analysis. In this chapter the raw data for every individual will first be transformed to the 6^{th} scale approximation using DWT with a Daubechies-4 wavelet.

The application of these methods follows in the results section.

3.1 Methods

Fundamental to all clustering techniques is the choice of distance between objects (Friedman et al., 2001). Similarity represents the degree of correspondence among objects. It must be determined between each of the observations to enable these to be compared to each other.

When similarity measures have been calculated, clusters can then be formed based on these. Typically a number of cluster solutions are formed, and then the final solution is selected from the set of possible solutions based on certain criteria. Clustering algorithms can be divided into two groups; hierarchical and non-hierarchical/partitional (Hair et al., 2006). Hierarchical algorithms move in a stepwise fashion to form an entire range of solutions, whereas partitional clustering algorithms find all clusters simultaneously as a partition of the data and do not impose a hierarchical structure. Hierarchical procedures can either be agglomerative, where each observation starts as its own cluster and are recursively joined to form one large cluster, or divisive where the procedure starts with all observations in one cluster and recursively divides it.

Given the vagueness of the term similarity, a cluster is a subjective entity whose

significance and interpretation requires domain knowledge. The researcher should have a strong conceptual basis to deal with issues such as why groups exist in the first place (Hair et al., 2006). To clarify what exactly the clustering will be based on, the measure of similarity that is used will be made explicit in the next section.

The hierarchical procedure will then be presented. This will be used in order to help in the detection of any outliers and also to identify a preliminary range for the number of clusters (K). Having removed the outliers, a non-hierarchical procedure will be applied using the range of K suggested by the hierarchical methods. Silhouette analysis will then be employed to determine the optimal value of K.

3.1.1 Similarity

The most commonly used measures of similarity are distance measures, they represent similarity as the proximity of observations to one another (Hair et al., 2006). These measures are actually a measure of dissimilarity, with larger values denoting less similarity. Of these distance measures, the most commonly used is the Euclidean distance (Anton & Rorres, 2010), which is what will be used here to cluster observations. The Euclidean distance is given by:

$$d = (\sum_{i=1}^{n}(|x_i - y_i|)^2)^{\frac{1}{2}} \tag{3.1}$$

Two individuals in the cohort, denoted X and Y, are represented by vectors of the transformed activity levels, denoted by $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ respectively. In this study the activity levels are aggregated into 1 minute intervals, of which there are 1440 such intervals in a day, therefore $n = 1440$ in Equation (3.1). To illustrate this, the first two individuals in the cohort, case IDs 8 and 138, are taken as X and Y respectively. The curves for both are shown in Figure 3.1, which also highlights the activity level for each at the midday interval.

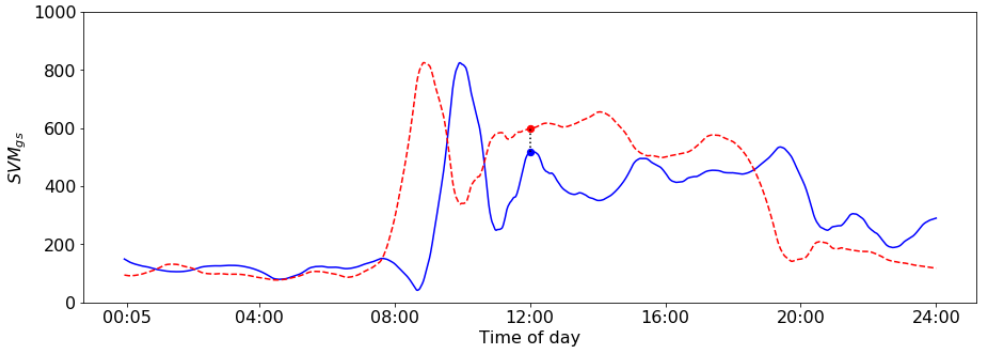


Figure 3.1: Similarity measure between IDs 8 and 138

The difference between x_{720} and y_{720} (where 720 represents the midday point), along with the difference between every other pair of points is summed. Thus Equation (3.1) gives:

$$d = (\sum_{i=1}^{1440}(|x_i - y_i|)^2)^{\frac{1}{2}} = 7542 \tag{3.2}$$

This is calculated between every pair of individuals and is used as the basis for joining individuals into clusters.

3.1.2 Hierarchical

Having decided how to join single observations into clusters, the next question is how to join clusters that have multiple members. Do we use a single member of the cluster for linkage, or use some composite to represent the cluster. The five most popular methods for doing so are single linkage, complete linkage, average linkage, centroid method and Ward's method (Hair et al., 2006).

Single linkage, or nearest neighbour, joins clusters based on the proximity of their two closest objects. Whereas complete, or farthest neighbour, joins clusters based on the proximity of their two furthest away objects. Average linkage, is the average of distances between all pairs of objects. Centroid method is based on the distance between the centroids of each cluster. A cluster centroid contains the averages for each variable based on its members. Ward's method (Ward Jr, 1963) looks at cluster analysis as an analysis of variance problem. The distance between two clusters is the sum of squares, at each stage the within cluster sum of squares is minimised. This technique tends to combine clusters with a small number of observations, and so is biased towards creating clusters with approximately the same number of observations. For the purposes of this study, only the single and complete linkage methods will be used.

An important characteristic of hierarchical procedures is that the results at an earlier stage are always nested within the results at a later stage, creating a similarity to a tree (Hair et al., 2006). As clusters are formed only by joining existing clusters, any member of a cluster can trace its membership in an unbroken path to its beginning as a single observation. This can be illustrated using a dendrogram. A dendrogram is a visualization in the form of a tree showing the order and distances of merges during the hierarchical clustering. To demonstrate a dendrogram, the 10 individuals are clustered with a complete linkage clustering method, as shown in Figure 3.2.

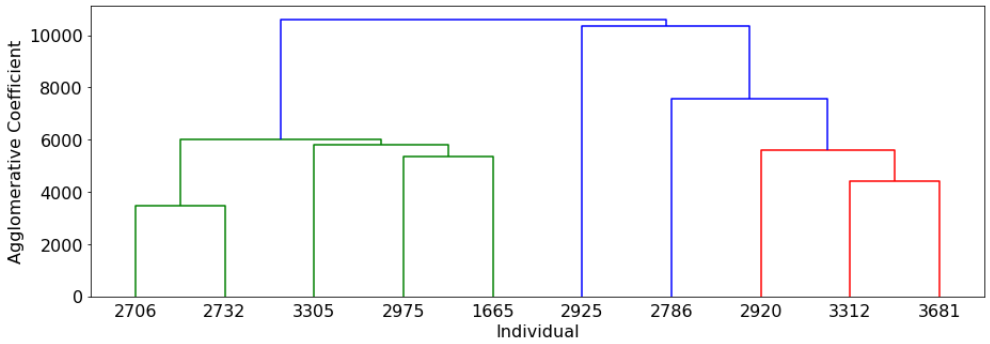


Figure 3.2: Dendrogram example

The agglomerative coefficents are the Euclidean distances as calculated in the previous section. Based on the agglomerative coefficients in this sub-sample, the pairs of individuals (2706, 2732) and (3312, 3681) are the first clusters to be formed. Methods like this result in an agglomerative schedule, which is the order in which clusters are merged, and can also be used to determine the number of clusters to be considered (Milligan, 1980).

In the search for structure, cluster analysis is sensitive to outliers (Hair et al., 2006). Outliers can either be truly atypical observations that are not representative of the population; they can represent small or insignificant segments within the population. Or they can be an under-sampling of actual groups in the population. There is no one way to detect outliers, instead a number of mechanisms are examined and collectively, can be used to form an opinion of which observations are deemed to be outliers. These include the examining of box plots, in addition to the agglomerative schedule. The single-linkage method is one of the simplest agglomerative hierarchical methods that is commonly used to detect an outlier. Using a dendrogram with a single linkage clustering method, individuals that are far removed from everyone else can be identified, this is illustrated for the same 10 individuals in Figure 3.3.

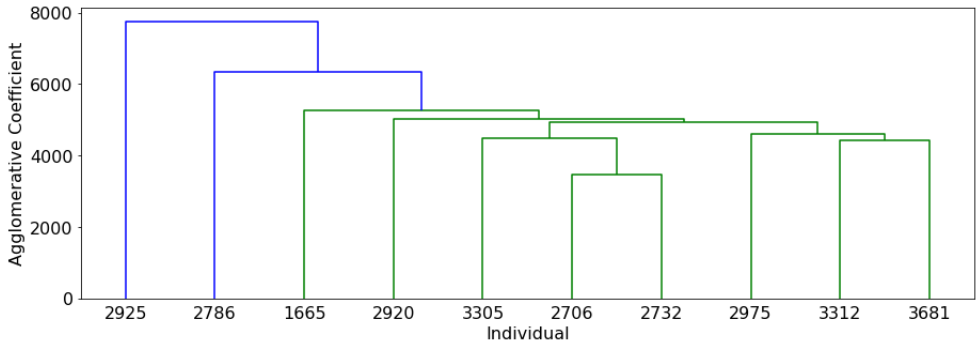


Figure 3.3: Dendrogram example for single linkage

What this illustrates is that ID: 2925 is the most dissimilar to all others in this

sub-sample, as the other nine individuals are joined into a single cluster before ID: 2925 is merged. When applied to the cohort as a whole, this should identify individuals that are dissimilar to everyone else and could potentially be outliers.

Once detected, these outliers can be removed from the data and further clustering can be performed. The solutions from hierarchical clustering are impacted by a common characteristic that once observations are joined in a cluster, they are never separated or reassigned in the clustering process (Hair et al., 2006). Non-hierarchical procedures have the advantage of being able to better optimize cluster solutions by reassigning observations until maximum homogeneity within clusters is achieved.

The number of clusters identified by the hierarchical procedure can be utilised in the non-hierarchical approach. The agglomerative coefficient can be plotted against the number of clusters to obtain a representation of the point at which the largest change in the coefficient occurs. This is known as the elbow method (Thorndike, 1953), and can be used to determine the number of clusters. An example of the elbow is presented in Figure 3.4.

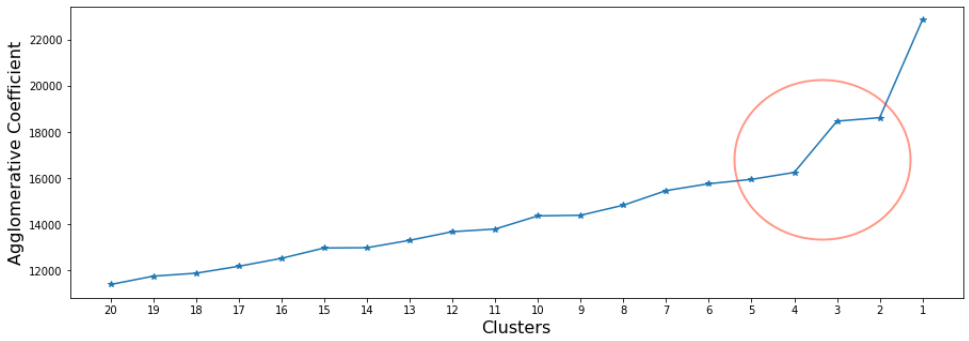


Figure 3.4: Elbow method example

The analogy for this method is that the plot looks like an arm, with the elbow then being the optimal number of clusters. The underlying idea for this method is that one should choose a number of clusters so that adding another cluster does not give better modelling of the data (Bholowalia & Kumar, 2014).

3.1.3 Non-Hierarchical

In contrast to hierarchical procedures, these methods do not follow an agglomerative or divisive schedule. Therefore they do not follow the tree-like process of the hierarchical methods and can not be represented using a dendrogram. The number of clusters is specified first and then objects are assigned into clusters. For example, a three cluster solution is not just the merger of two clusters from the four cluster solution, as was the case for

hierarchical methods.

K-means (Romesburg, 2004; Hair et al., 2006) is one of the most widely used clustering algorithms. Its simplicity, efficiency and ease of implementation are some of the reasons for its popularity. Given a representation of n objects, it finds K groups based on some measure of similarity. It finds a partition such that the squared error between the empirical mean of a cluster and the points in a cluster is minimized.

The algorithm requires three parameters: number of clusters K, a distance metric, and an initial set of cluster centres. The value of K will be identified from the hierarchical clustering. A typical distance metric used is the Euclidean metric for computing the distance between points and cluster centres. Different initializations of cluster centres can lead to different final centres; to avoid this convergence to local minima, the algorithm can be run multiple times with different initializations.

The cluster centroids from the hierarchical solution can be used as the cluster seeds, or they can be randomised. K-means typically uses sequential threshold, parallel threshold or optimization for assigning individuals to clusters. Sequential threshold starts by selecting one cluster seed and includes all observations within a prespecified distance. When all observations within the prespecified distance are included, a second cluster seed is selected and so on. Once an observation is clustered with a seed, it is no longer considered for subsequent seeds. Parallel threshold selects several cluster seeds and assigns observations within the threshold distance to the nearest seed. Some observations can remain unclustered if they are outside the prespecified distance. Both these methods have the reassignment issue seen with hierarchical procedures. The optimization method is similar to the other two but allows for reassignment. If an observation becomes closer to another cluster, different from its original cluster, it will switch cluster. Given this property, this method will be used as the reassignment of observations is desirable.

The reassignment of observations works in the following manner. Using an initial seed for the centroids, each individual is assigned to a cluster. Once they are all assigned, the centroids are recalculated, and any individual that is now closer to a centroid that is not its own is reassigned to the cluster it is closer to. This process continues iteratively until a stable solution is reached.

One of the hardest tasks in K-means clustering is to determine the appropriate number of clusters K. The elbow method, discussed earlier, is one example of a visual aid that can be utilised to help determine this number. Silhouette analysis (Rousseeuw, 1987) is another, and it used to study the separation distance between clusters. It measures how close each point in one cluster is to

points in the neighbouring clusters. The silhouette coefficient, s, for a given individual is defined as:

$$s = \frac{b - a}{max(a, b)} \tag{3.3}$$

where a is the mean distance between an individual and all others in the same cluster, and b is the mean distance between an individual and all other points in the next nearest cluster.

The score is bounded between [-1, 1], where a score -1 indicates incorrect clustering and that an individual has been misclassified. A score of +1 indicates highly dense clustering where individuals in a cluster are close to each other and far away from other clusters. Scores around zero indicate overlapping clusters. The score is higher when clusters are dense and well separated, which is desirable for a cluster.

The elbow method will be used in the hierarchical procedure to determine a range of potential values for K. Then the elbow method and silhouette analysis will be used in the non-hierarchical procedure to help determine which of these is the optimum K. The elbow method here will look at the percentage of variance explained as a function of the number of clusters. K-means will be implemented for a range of values of K, and the sum of squared distances of samples to their nearest cluster centre will be calculated. As K increases, the sum of squared distances tends to zero. If K was set to the maximum value, namely the number of individuals, every individual would be in their own cluster and so the the sum of squared distance will be zero. The first clusters will add information but at some point the marginal gain will drop dramatically and gives the elbow to the graph.

When implementing K-means, to make sure that the results are not influenced by the order of the individuals in the dataset, they will be randomized before the execution. To ensure stability and robustness of the results, the algorithm was run 10000 times with different centroid seeds. The final results will be the output in terms of the lowest inertia (Rousseeuw, 1987). The inertia is the sum of squared distances of observations to their closest centre.

3.2 Results

3.2.1 Similarity

The distance measure was calculated between every pair of observations. Table 3.1 gives a snippet of these values for 5 individuals.

	ID				
ID	2706	3312	2920	2925	2732
2706	-	-	-	-	-
3312	5366.0	-	-	-	-
2920	7996.0	5591.0	-	-	-
2925	9498.0	8162.0	9071.0	-	-
2732	3468.0	6481.0	9478.0	10598.0	-

Table 3.1: Similarity measure: Euclidean distance

The distance metric between IDs 8 and 138 was presented in the Methods section. Using an agglomerative approach, where each individual starts in their own cluster, the lowest distance was chosen and this pair forms the first cluster. The first three clusters that were formed using an agglomerative approach are shown in Figure 3.5.

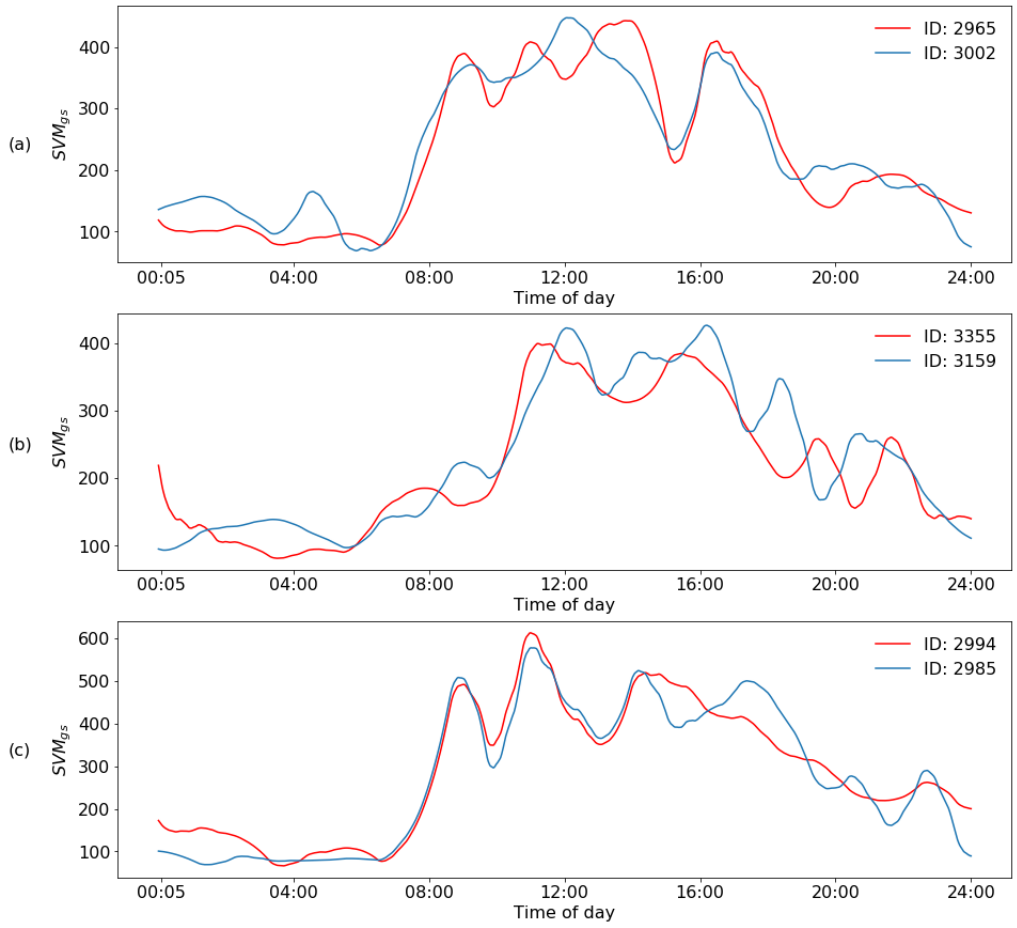


Figure 3.5: First three clusters with agglomerative approach. (a) 2965 & 3002 (1520), (b) 3355 & 3159 (1755), (c) 2994 & 2985 (1757)

The agglomerative coefficients for the first three clusters formed are given in the

brackets.

3.2.2 Outlier Detection

Using a dendrogram with a single linkage clustering method, individuals that are far removed from everyone else can be identified. Figure 3.6(a) shows this, while Figure 3.6(b) then focuses on the left most part of the graph and truncates the rest.

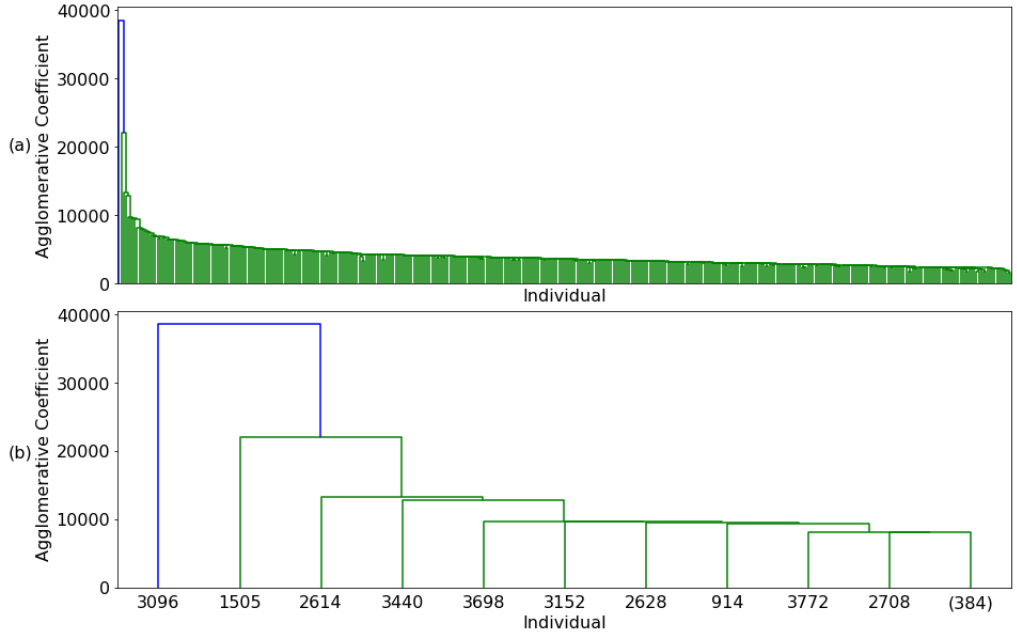


Figure 3.6: Hierarchical clusters formed using single linkage. (a) Full, (b) Truncated

What this illustrates is that 384 individuals, the number in brackets in Figure 3.6(b), of the 394 in total, are joined into a single cluster before any cluster is formed with the other 10 individuals shown here. The large agglomerative coefficients in the graph illustrates that these 10 are largely dissimilar to everyone else. This suggests the potential for these being outliers and requires further investigation. The agglomerative coefficients for these 10 individuals are given in Table 3.2.

ID	Agglomerative Coefficient
2708	8054
3772	8115
914	9398
2628	9577
3152	9599
3698	9687
3440	12803
2614	13345
1505	21996
3096	38571

Table 3.2: Potential outliers

These represent the minimum distance between these IDs and any other individual, and further clarify, what is illustrated in Figure 3.6, that these IDs are far removed from everyone else. Looking at the top ten individuals in terms of the raw total activity (TA) over the day may provide more information about these individuals, these values are given in Table 3.3. Those highlighted in red also appear in the potential outliers table (Table 3.2).

ID	Total PA
3096	2298792
2614	1164623
1505	1161222
3440	1068952
3698	995812
2628	986669
1862	857701
3792	821708
3306	813537
2633	809630

Table 3.3: Top ten individuals by total activity

As evident in Table 3.3, ID 3096 is extremely active. The boxplot in Figure 3.7 further illustrates the disparity of these six individuals from the rest of the cohort.

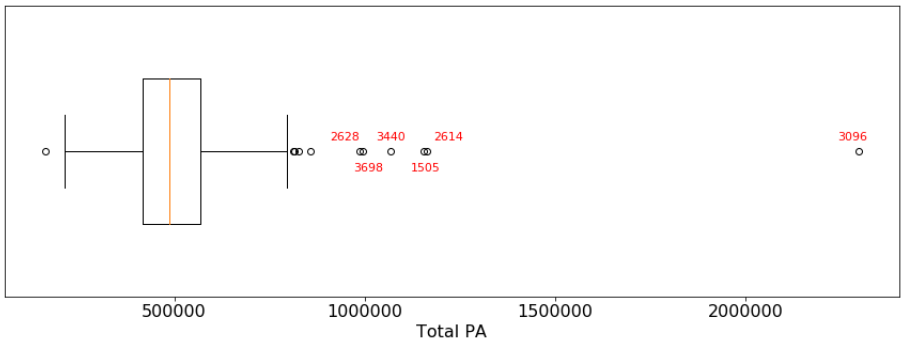


Figure 3.7: Boxplot of total physical activity

This means that they may be outliers and are not simply a smaller cluster. They are not representative of the cohort and so will be removed for the remainder of the cluster analysis to avoid distorting the actual structure of the cohort.

3.2.3 Determining K

Having removed the outliers, other linkage methods can be explored to determine a range for the number of clusters (K), Figure 3.8(a) shows the hierarchical clusters formed using the complete linkage method. Figure 3.8(b) shows a truncated version of the same dendrogram which displays the final few cluster mergers, in addition to the number of individuals in each cluster along the x-axis.

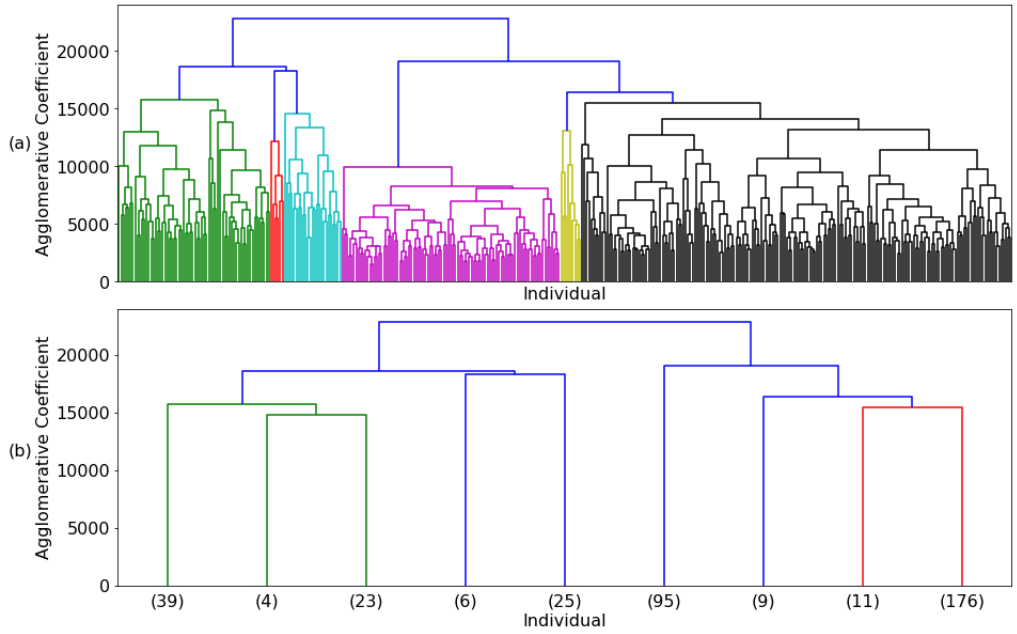


Figure 3.8: Hierarchical clusters: Complete linkage

To further investigate what values of K are appropriate, the agglomerative coefficients will be examined. Table 3.4 shows these values for the last 10 cluster mergers using the complete linkage method.

Clusters (i)	AC_i	AC_{i-1}	% Change (i to $i-1$)
10	14065	14551	3.46
9	14551	14821	1.86
8	14821	15464	4.34
7	15464	15751	1.86
6	15751	16379	3.99
5	16379	18285	11.64
4	18285	18613	1.79
3	18613	19076	2.49
2	19076	22845	19.76
1	-	-	-

Table 3.4: Complete linkage: Agglomerative schedule

Small coefficients represent fairly homogenous clusters, whereas large coefficients or a large percentage change in the coefficients indicates heterogenous clusters. The largest percentage increase occurs when going from two clusters to one, which makes sense as these are the most dissimilar, as evident in Figure 3.8(a) where the highest agglomerative coefficient is when the last two clusters merge. The next largest increase is from five clusters to four, highlighted in Table 3.4, which indicates that two dissimilar groups were merged. This suggests that five clusters is a potential value for K. To further illustrate the changes in agglomerative coefficient, it is plotted against the number of clusters in Figure 3.9.

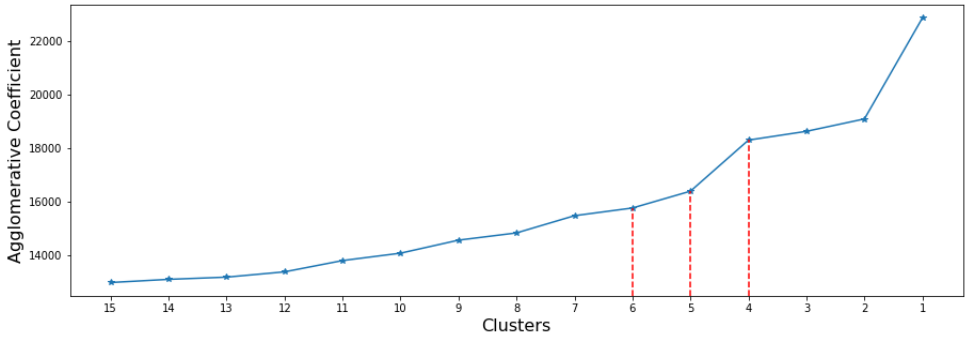


Figure 3.9: Complete linkage: Agglomerative coefficient vs. Number of clusters

Figure 3.9 suggests that K lies between four and six. Therefore based on the hierarchical clustering, K-means will be analysed for K equal to four, five and six.

3.2.4 K-means

K-means will be implemented for each of the specified number of clusters using the optimized threshold method. The centroids, which represent the average profile in the group, for each cluster are shown in Figure 3.10. The number of constituents for each cluster are shown in brackets in the legend.

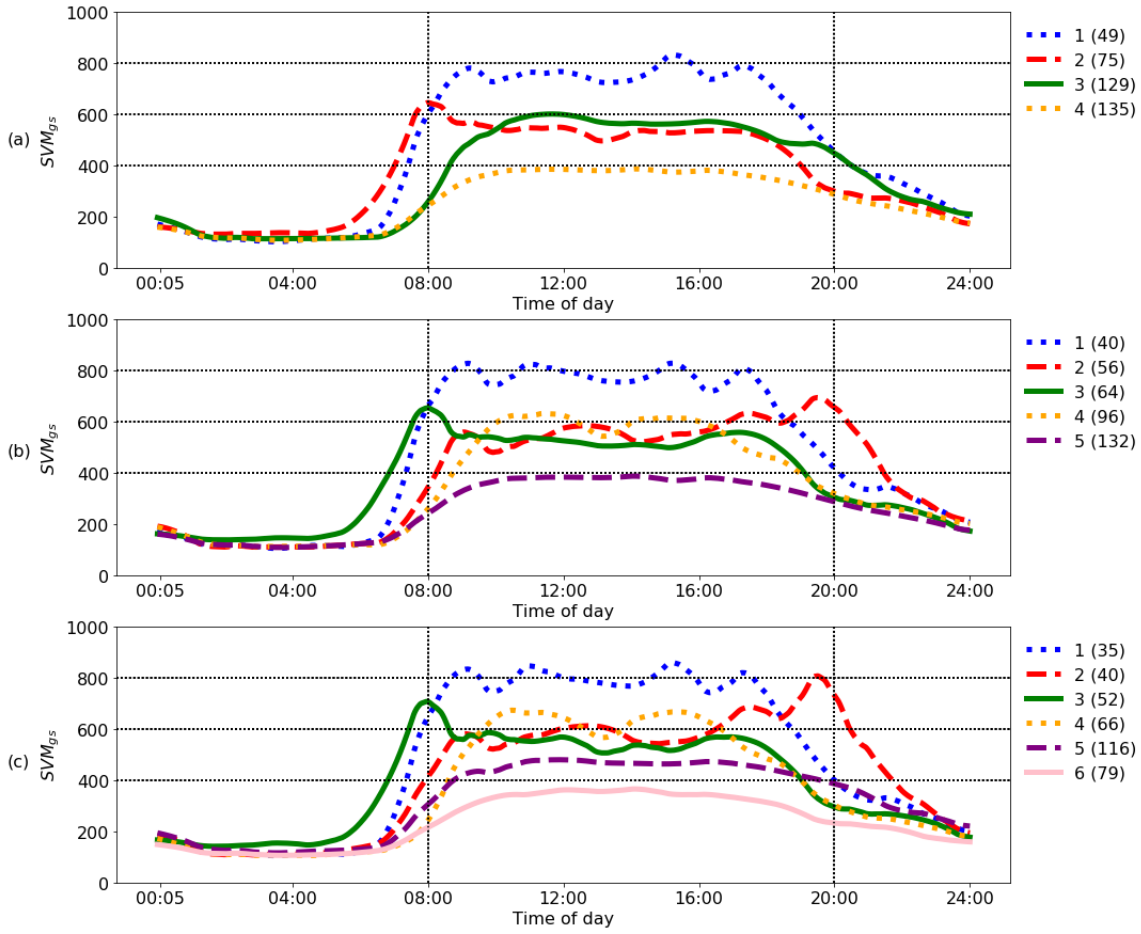


Figure 3.10: Centroids for K-Means clustering. (a) K=4, (b) K=5, (c) K=6

The dashed lines were added to the graphs to help distinguish characteristics of the different solutions. The horizontal lines at 400, 600 and 800 can be loosely thought of low, moderate and high activity levels. While the vertical lines at 8a.m. and 8p.m. can be thought of as the beginning and end of the day. The clusters are colour coded based on the centroid's max activity level over the day, using the decreasing colour series (blue, red, green, orange, purple, pink). Therefore, the top cluster is always blue in Figure 3.10. The alternating line styles are to help further distinguish the centroids. The centroids for the four cluster solution are clear-cut. Cluster 1 (blue) is the high activity group, whereas cluster 4 (orange) is the low activity group, and their centroids in Figure 3.10(a)

are well separated. The two moderate activity clusters are distinguished by their morning activity, with cluster 2 (red) being more active before 8a.m. compared to cluster 3 (green). To get an idea of the make-up of these clusters, Figure 3.11 shows the four centroids and 20 randomly selected constituents from each cluster.

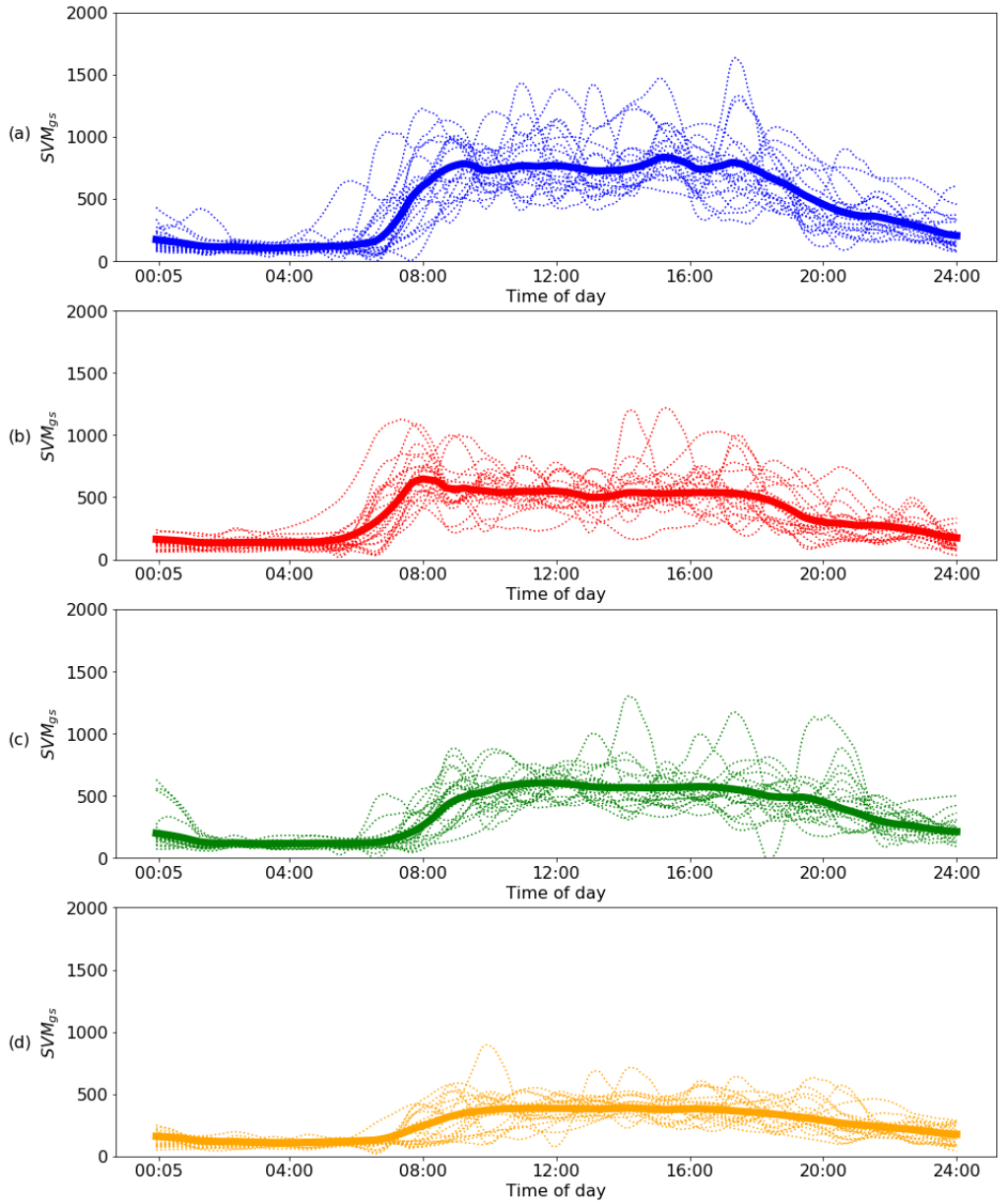


Figure 3.11: Four cluster solution. (a) 1, (b) 2, (c) 3, (d) 4

Individuals in cluster 1 (blue) are quite active compared to the other clusters. Cluster 2 (red) can be characterised as early risers, or morning larks, and prefer getting up and going to bed early, and are at peak performance early in the day. This is evident by the peaks in the activity profiles before 8a.m. The opposite of morning larks are night owls, who like sleeping in and staying up late, and do

not perform well until the afternoon or evening.

The psychological term for describing these characteristics is chronotype. It describes an individual's propensity for sleep and activity at particular times during a 24 hour period. It is the expression of circadian rhythmicity in an individual, and three categories of chronotype are defined: morning types (M-types), evening types (E-types), and neither types (N-types). M-types generally wake up and go to bed early (Taillard, Philip, Chastang, & Bioulac, 2004) and have their best performances in the first part of the day, whereas E-types go to bed and wake up late and have their peak performances in the evening (J. Horne, Brass, & Petitt, 1980). These M-types have already been characterised in the four cluster solution, cluster 2 (red). However, there is no E-type cluster evident which therefore warrants further investigation into the five cluster solution.

In the five cluster solution, Figure 3.10(b), cluster 3 (green) represents the early risers (M-types), and cluster 2 (red) represents the night owls (E-types). Clusters 1, 4 and 5 represent sub-divides of the N-types, which can be thought of as high, moderate and low activity clusters respectively. In the six cluster solution, Figure 3.10(c), these early risers and night owls clusters still exist, and the N-types are now sub-divided into four activity levels.

Through the profiling of these clusters, subgroups of early risers and night owls were identified and characterised. These subgroups exist for both K=5 and K=6, therefore K=4 will no longer be considered. The choice of K is now a question of whether to divide the N-types into three or four subgroups. These subgroups are illustrated in Figure 3.12.

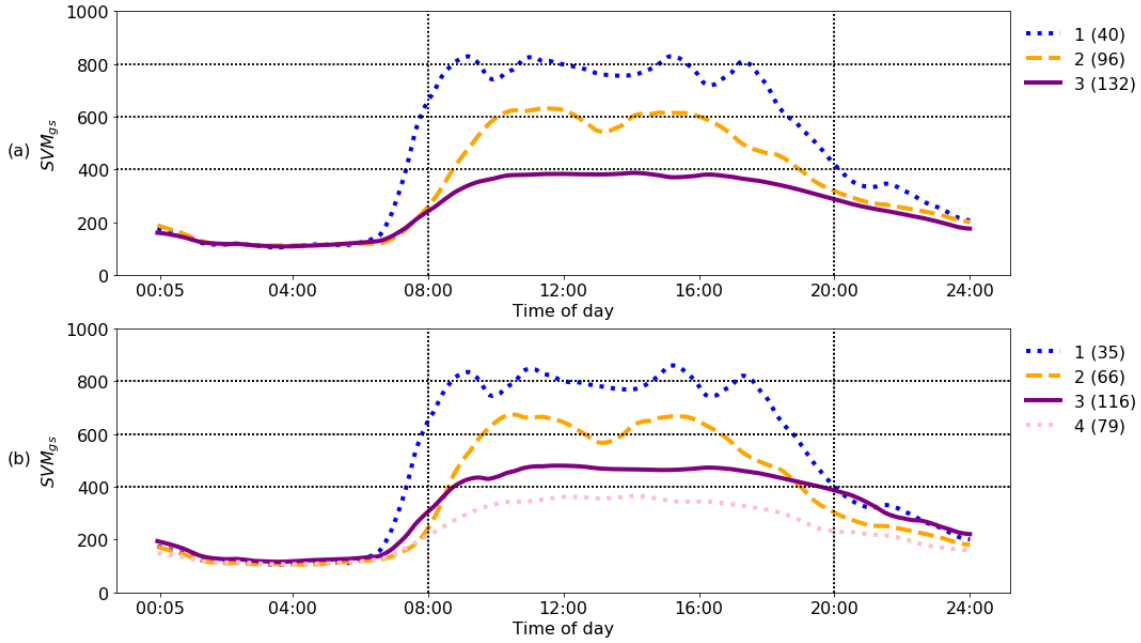


Figure 3.12: N-types. (a) k=5, (b) k=6

Conceptually the three levels in Figure 3.12 are easier to interpret as high, moderate and low. To determine which value of K is optimum, visual aids are utilised. The elbow method, shows the sums of squared distances for increasing values of K in Figure 3.13.

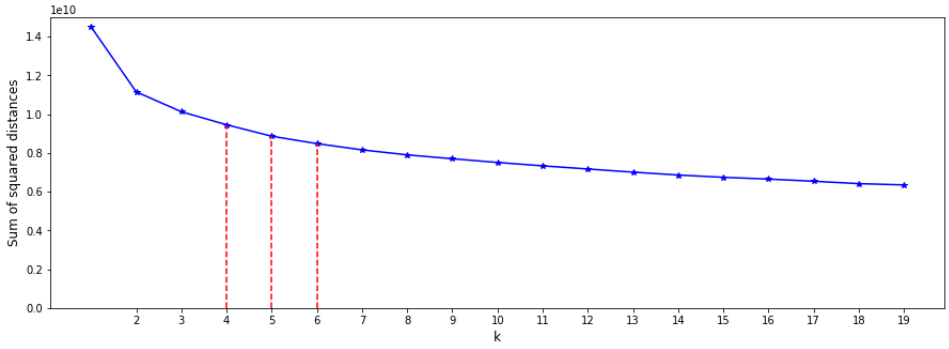


Figure 3.13: Sum of squared distances vs K

The elbow method suggests that the optimal number of clusters is five, however, this elbow can not always be unambiguously identified. Another visual aid to help in determining the optimal number of clusters is silhouette analysis, the silhouette coefficients for each of the cluster's constituents are shown in Figure 3.14. The thickness (along the y-axis) of the silhouette plot can be used to visualize the size of the cluster. For example, cluster 5 (purple) in Figure 3.14(a) is the largest cluster.

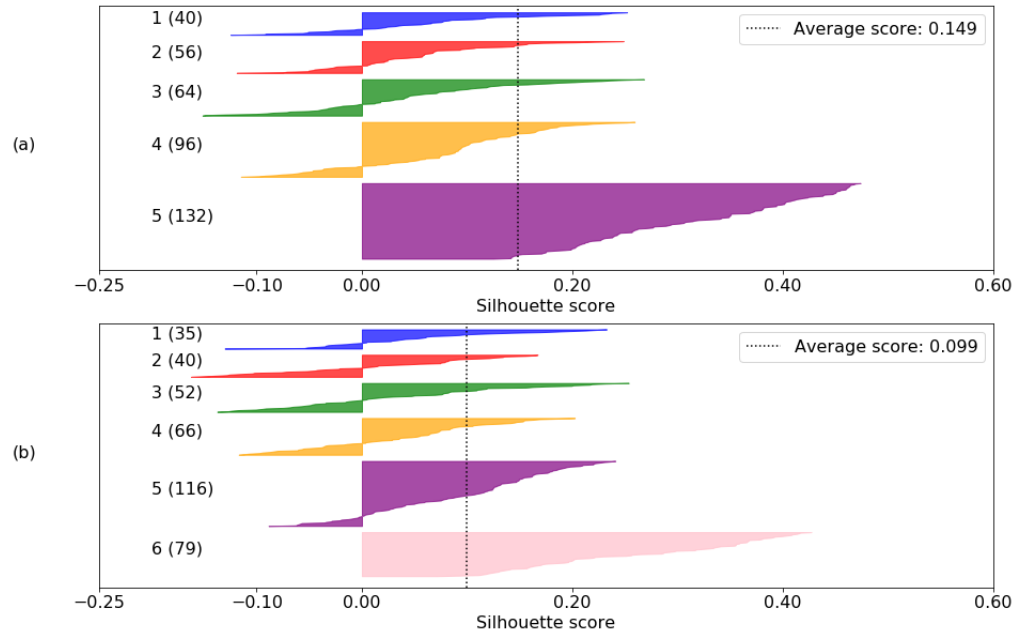


Figure 3.14: Silhouette analysis. (a) K=5 (b) K=6

None of the silhouette scores are close to +1, meaning that the clusters are not going to be well separated. This was expected given the dispersion of individuals within clusters, as seen in Figure 3.11. Given the density of the individual curves, well separated clusters was not achievable. Cluster 5 (purple) in Figure 3.14(a), and cluster 6 (pink) in Figure 3.14(b), both represent the lowest activity N-type clusters, and neither have any negative silhouette scores. This tells us that the lowest activity cluster remains well separated from the other individuals for both K=5 and K=6.

To illustrate the negative silhouette scores, the centroids of the five cluster solution will be plotted along with the individual in each cluster with the lowest silhouette score, as shown in Figure 3.15.

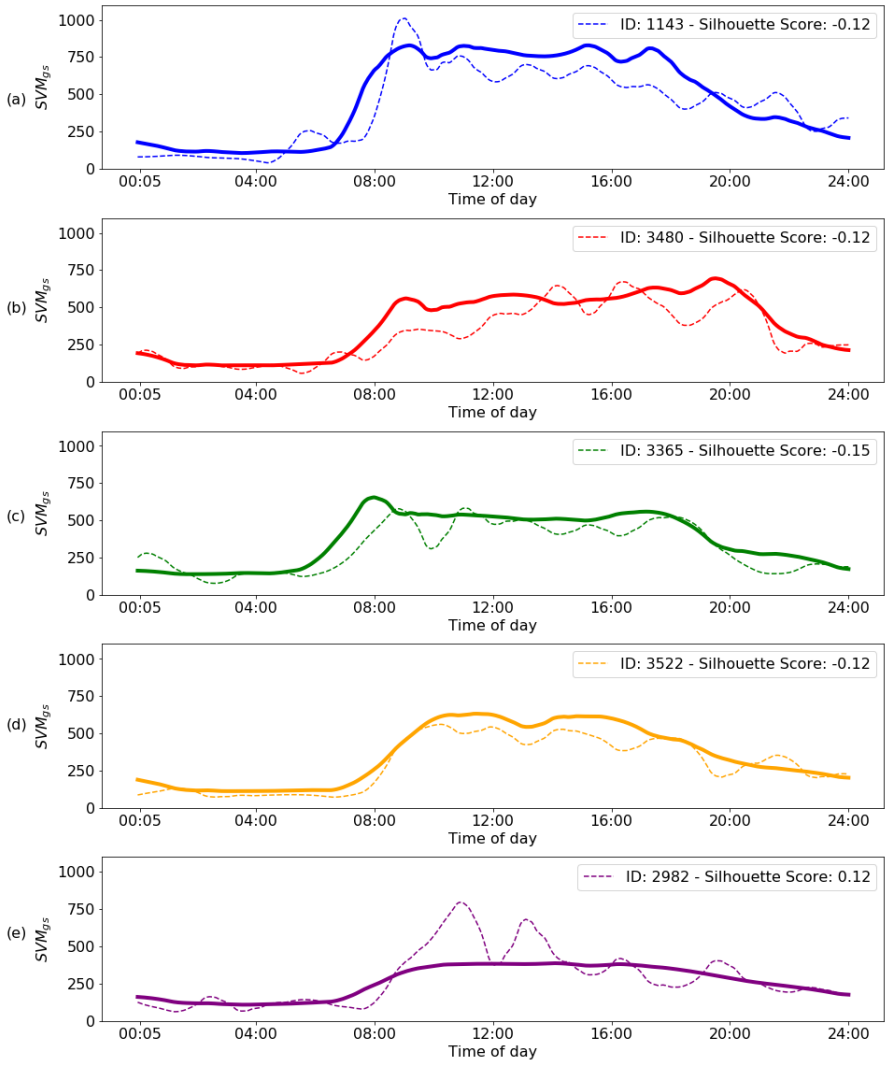


Figure 3.15: Lowest silhouette score (a) 1: N-type (High), (b) 2: E-type, (c) 3: M-type, (d) 4: N-type (Moderate) , (e) 5: N-type (Low)

The individual (ID 3365) in Figure 3.15(c) has the lowest silhouette score, and this will be used to illustrate the calculation. First the mean distance between ID 3365 and all others in the same cluster is calculated. Then the mean distance between ID 3365 and all others in the other clusters is calculated. The next nearest cluster is defined as the lowest of these values. These values are shown in Table 3.5.

Cluster	Mean Distances
1: N-type (High)	10226
2: E-type	7290
3: M-type	6064
4: N-type (Moderate)	6027
5: N-type (Low)	5145

Table 3.5: Silhouette mean distances for an individual to all other clusters

These values are then plugged into the silhouette score formula, Equation (3.3) to give the following:

$$s = \frac{b-a}{max(a,b)} = \frac{5145 - 6064}{max(5145, 6064)} \approx -0.15 \tag{3.4}$$

ID 3365 is closer on average to the individuals in cluster 5 N-type (Low) but they possess the downward trend which is characteristic of cluster 3 M-type and so can be considered to be clustered correctly.

The average silhouette score over all points in a cluster is a measure of how tightly grouped all the points in the cluster are. Therefore, the average over all the data is a measure of how appropriately the data has been clustered. Thus, to choose between K=5 and K=6, the average value should be as close to 1 as possible. As a value of +1 is considered ideal, then the higher the value, the better the cluster configuration. In which case K=5 would be chosen as optimum.

The clustering in this chapter was performed purely using distance metrics. In the next chapter, functional principal component analysis is performed, which focuses on features rather the data. Having extracted these features, cluster analysis will be performed again on these features. The results of this will then be compared to those that were obtained in this chapter. Having decided on five as the optimal number for K, this will be used when looking to cluster using the dominant patterns in the data. This is the subject of the next chapter.

Chapter 4 - Functional Principal Component Analysis

In functional data analysis (Ramsay, 2005) (FDA), the individual datum is a whole function bounded on a common interval, rather than concentrating on the observed values at particular points in the interval. The idea is that a function f is a single object, which may itself vary and is thought of as a point in a functional space.

FDA extends existing methodologies and theories from multivariate data analysis. Indeed, functional principal component analysis (Ramsay, 2005; Hall, Müller, Wang, et al., 2006) (FPCA), is an extension of the classical principal component analysis (PCA) (Pearson, 1901), which will be detailed in the methods section. PCA was one of the first multivariate data analysis methods to be adapted to functional data (Dauxois, Pousse, & Romain, 1982). FPCA is a method for investigating the dominant modes of variation of functional data. In our cohort, it has already been observed that the data is functional.

Each individual has repeated measurements of activity counts over the course of a day. As seen in Chapter 2, smoothing procedures can yield a functional representation of a finite set of observations. Each individual's data was smoothed to reveal the underlying functional structure and reduce the effects of noise. Now each individual can be represented by a curve (or function), based on his/her observed data. The statistical analysis of n such curves is commonly termed FDA (Ramsay, 2005).

In FDA, the data can be explored to see the features that characterise typical functions. FPCA decomposes functional data into population level basis functions and subject-specific scores (Ramsay, 2005), and is used to investigate the dominant patterns in the data.

The point of this analysis is to find several "eigen-time-series", that would describe the typical shape of the curves in the cohort. Eigenvectors and eigenvalues are used to describe the variance of a dataset, and will be detailed further in the Methods section. Each individual's curve can be written as a weighted sum of eigencurves. Having deduced the functional principal components, the related scores will be clustered and compared to the results from Chapter 3.

4.1 Methods

In order to review FPCA, it is necessary to first look at PCA, and how it is used to reduce dimensionality for multivariate data. The goal of PCA is to find

the sequence of orthogonal components that most efficiently explains the variance of the observations. PCA finds a lower-dimensional representation of the data, while preserving the maximum amount of information from the original variables.

The first step in PCA is to centre the data in order to simplify the notation and computations. This is done by subtracting the mean for each of the data dimensions, or variables. Each dimension in our case is the activity at each of the time points t, where $t = 1, 2, 3, ..., 1440$, from the smoothed curves such that there are 1440 dimensions (p). The mean subtracted is the average across each dimension, or the average activity across individuals at a given time. This produces a dataset whose mean is zero. The data can be represented by a matrix:

$$X_0 = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,p} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,p} \end{pmatrix}$$

where n is the size of the sample (388) and p is the number of variables (1440).

To illustrate the effect of this mean centring, the plots for the first two individuals in the cohort (IDs 8 and 138), will be shown. A snippet of the values for IDs 8 and 138, along with means for each variable is shown in Table 4.1.

				t				
ID	1	2	3	4	5	...	1439	1440
8	150	148	147	146	145	...	290	290
138	94	94	94	93	93	...	118	118
...	...	...	...	...	...	...	...	...
Mean	178	178	177	177	176	...	194	193

Table 4.1: Variable means

The effect of this mean centring for these two individuals is shown in Figure 4.1. Each individual's curve can be reconstructed by adding the mean curve ($\mu(t)$) and its difference from $\mu(t)$.

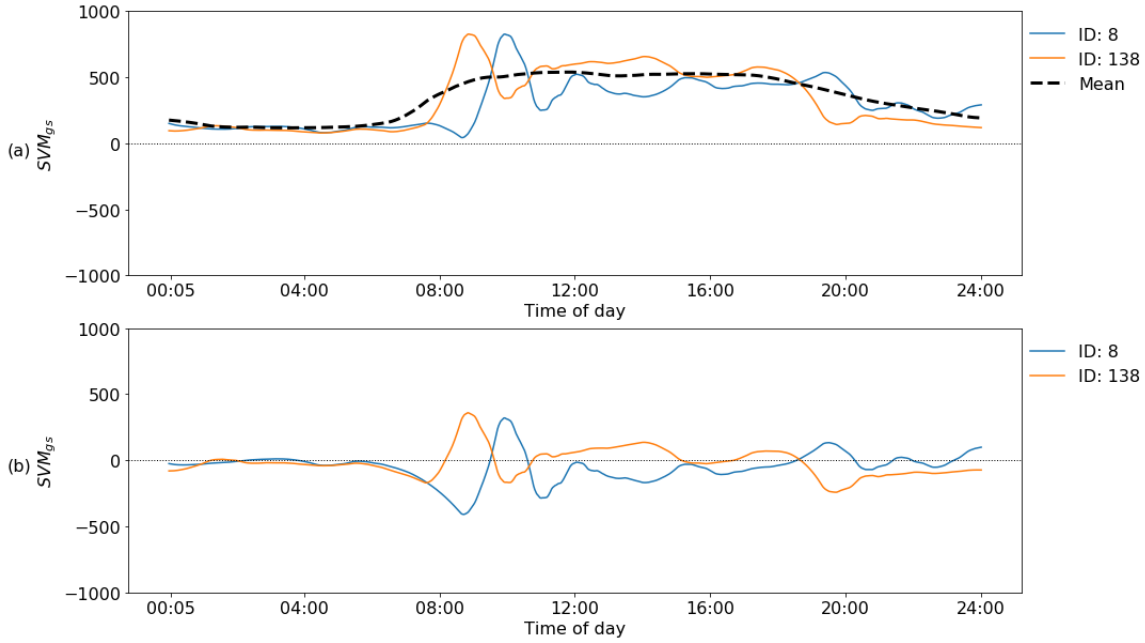


Figure 4.1: Profiles for IDs 1 and 8 (a) Normal, (b) Mean centred

Next the covariance matrix needs to be calculated. The data has 1440 dimensions, therefore, the covariance matrix will be 1440 x 1440. To illustrate, the first two of these dimensions, namely t={1, 2}, are taken to investigate how they move together. The covariance is given by:

$$cov(X_1, X_2) = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X_1})(X_{2i} - \bar{X_2})}{(n-1)} \tag{4.1}$$

where the vector X_1 will be the values for all individuals at time $t = 1$, and similarly X_2 is the values at $t = 2$. The covariance matrix is thus:

$$cov = \begin{pmatrix} 8904 & 8852 \\ 8852 & 8803 \end{pmatrix}$$

Since the covariance matrix is square, the eigenvectors and eigenvalues can be calculated. An eigenvector (v), is a vector which does not change direction in a transformation. A matrix is another name for a transformation, so if we label our covariance matrix A, then eigenvectors (v) and eigenvalues (λ) are such that they satisfy Equation (4.2):

$$Av = \lambda v \tag{4.2}$$

These eigenvalues and vectors are calculated and shown below:

$$eigenvalues = \begin{pmatrix} 17706 \\ 2 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} 0.7091 & -0.7051 \\ 0.7051 & 0.7091 \end{pmatrix}$$

To demonstrate what these eigenvectors represent, the two vectors X_1 and X_2 are plotted against each other, with the eigenvectors superimposed on the plot. This is shown in Figure 4.2

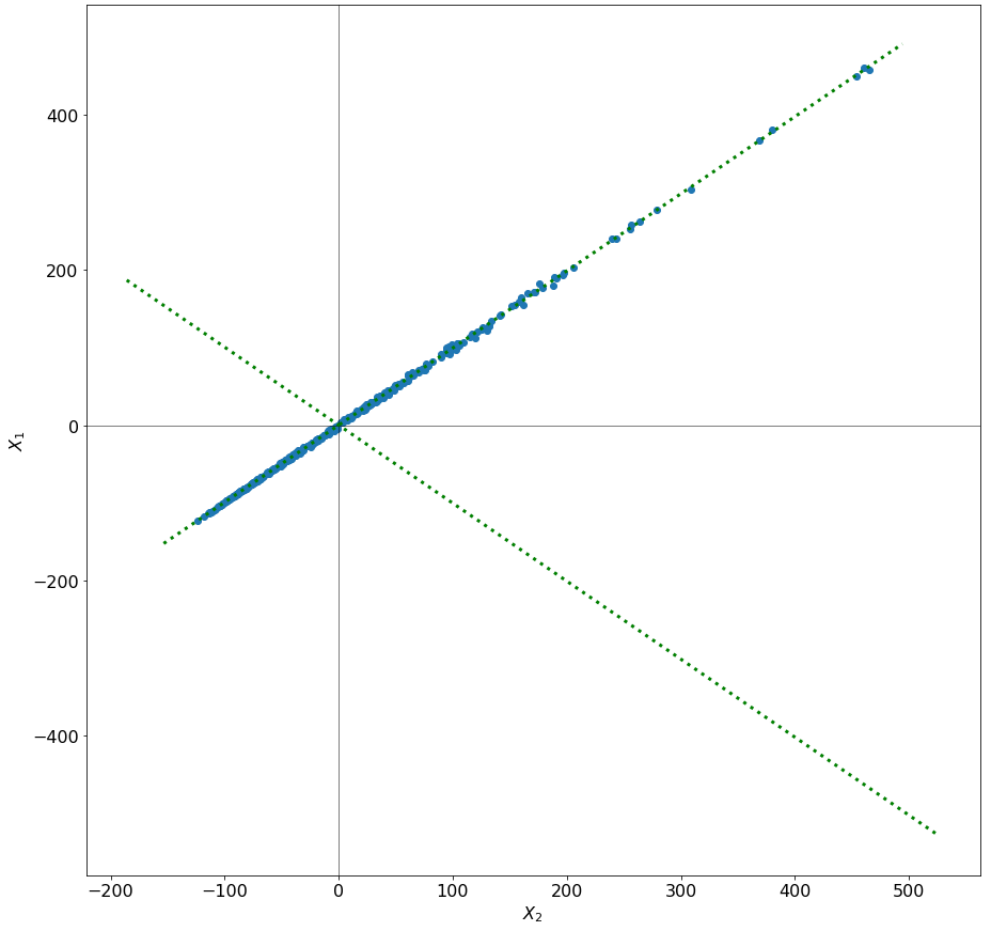


Figure 4.2: Scatterplot of X_1 vs X_2 with eigenvectors

The solid blue dots represent the values of X_1 and X_2, and the dashed green lines are the eigenvectors. As evident from Figure 4.2, the two variables increase together. The eigenvectors are perpendicular to each other and they also provide us information about the patterns in the data. One of the eigenvectors appears to fit the data quite well, similar to fitting a regression line. The second eigenvector describes very little variation in the data. By taking the eigenvectors of the covariance matrix, lines that characterise the data have been extracted.

The eigenvalues that were calculated are quite different in size. The size of the eigenvalue indicates how much variance can be explained by its associated eigenvector. Therefore, the eigenvector with the highest eigenvalue is the principle component of the dataset. In general, for our dataset of 1440

dimensions, the eigenvectors can be ordered by eigenvalue, from highest to lowest. By ignoring the components with lower eigenvalues, the dimensionality of the data is reduced without losing too much information.

For example, take 10 as the number of components to be kept. Using these 10 components, the original data is then transformed into a new dataset where each individual can be described solely in terms of these components. These principal components can be thought of as a set of basis functions, which account for as much variation as possible at each stage. Every individual can then be represented with 10 weights, one for each of these principal components, rather than an individual having 1440 observed variables. The original curve can be reconstructed by adding the mean to the linear combination of weights and basis functions:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} c_{ik}\phi_k(t) \tag{4.3}$$

where $\mu(t)$ is the mean, $\phi_k(t)$ the set of orthonormal basis functions and c_{ik} are the subject specific scores.

The computation of PCA runs into serious difficulties in analysing functional data because of the "curse of dimensionality" (Bellman, 2015). The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high dimensional spaces that do not occur in lower dimensions. The huge number of correlated dimensions tends to increase the complexity of the model. Also, the higher dimensions used, the sparser the data becomes.

Even if the geometric properties of PCA remain valid, the sample covariance matrix is sometimes a poor estimate of the population covariance matrix. FPCA provides a more informative way of examining the sample covariance than PCA.

The main idea of the extension from PCA to FPCA is to replace vectors with functions. Now each individual is described by one continuous function rather than 1440 discrete observations. Eigenfunctions are deduced instead of eigenvectors. FPCA finds the set of orthogonal functional principal components (FPC) that maximize the variance along each component. It finds the first FPC, $\phi_1(t)$, for which the variance of the principal component scores:

$$\beta_1 = \int_{t_1}^{t_p} \phi_1(t) f(t) dt \tag{4.4}$$

is maximized subject to $||\phi_1^2(t)|| = \int_{t_1}^{t_p} \phi_1^2(t)dt = 1$. Where β_1 is the set of first principal component scores with mean zero and $f(t)$ is the set of functional curves,

which are supported on the range $[t_1, t_p]$. This can be compared to the discrete equivalent for PCA, given by:

$$\beta_1 = \sum_{t=1}^{p} \phi_1(t) x_t \tag{4.5}$$

Successive FPCs are then found iteratively by subtracting the first principal component from each $f(t)$, then using Equation (4.4) with this new set of functions, with the additional constraint that,

$$\int \phi_i(t)\phi_j(t)dt = 0, \forall i \neq j \tag{4.6}$$

This constraint ensures the perpendicularity of the FPCs. Approaches to FPCA vary depending the sparsity or density of the data. These terms describe the percentage of the dataset entries that are populated or not. The sum of the sparsity and density should equal 100%. For sparse data, Yao et al. (2005) proposed Principal Analysis by Conditional Expectation (PACE). PACE aims to estimate eigenfunctions and eigenvalues of the covariance surface with irregularly spaced longitudinal data (Yao, Müller, & Wang, 2005). The main idea of PACE is to first formulate a raw covariance using pooled sparse longitudinal measurements and then apply a two-dimensional local polynomial smoother to estimate the covariance. Smoothing based on the pooled raw data has the effect of borrowing strength from all data.

In our dataset, for all 388 individuals there was 1440 equally spaced activity observations with no missing entries. Therefore, our dataset is considered dense and procedures to deal with sparse data will be not considered.

The procedure for calculating the eigen components and associated FPC scores can be summarised as follows.

1. Calculate the cross-sectional mean μ. A snippet of this calculation was seen in Table 4.1.

2. Calculate the cross-sectional covariance surface.

3. Perform eigen analysis on the covariance to estimate the eigenfunctions ϕ and eigenvalues λ.

4. Use numerical integration to estimate the corresponding scores β using Equation (4.4).

The R package fdapace will be used for this analysis. The package can be used for both sparse and dense data. Its working assumption is that a dataset is treated as sparse if it has, on average, less than 20 potentially irregularly sampled measurements per individual.

4.2 Results

The first four FPCs or eigenfunctions are shown in Figure 4.3.

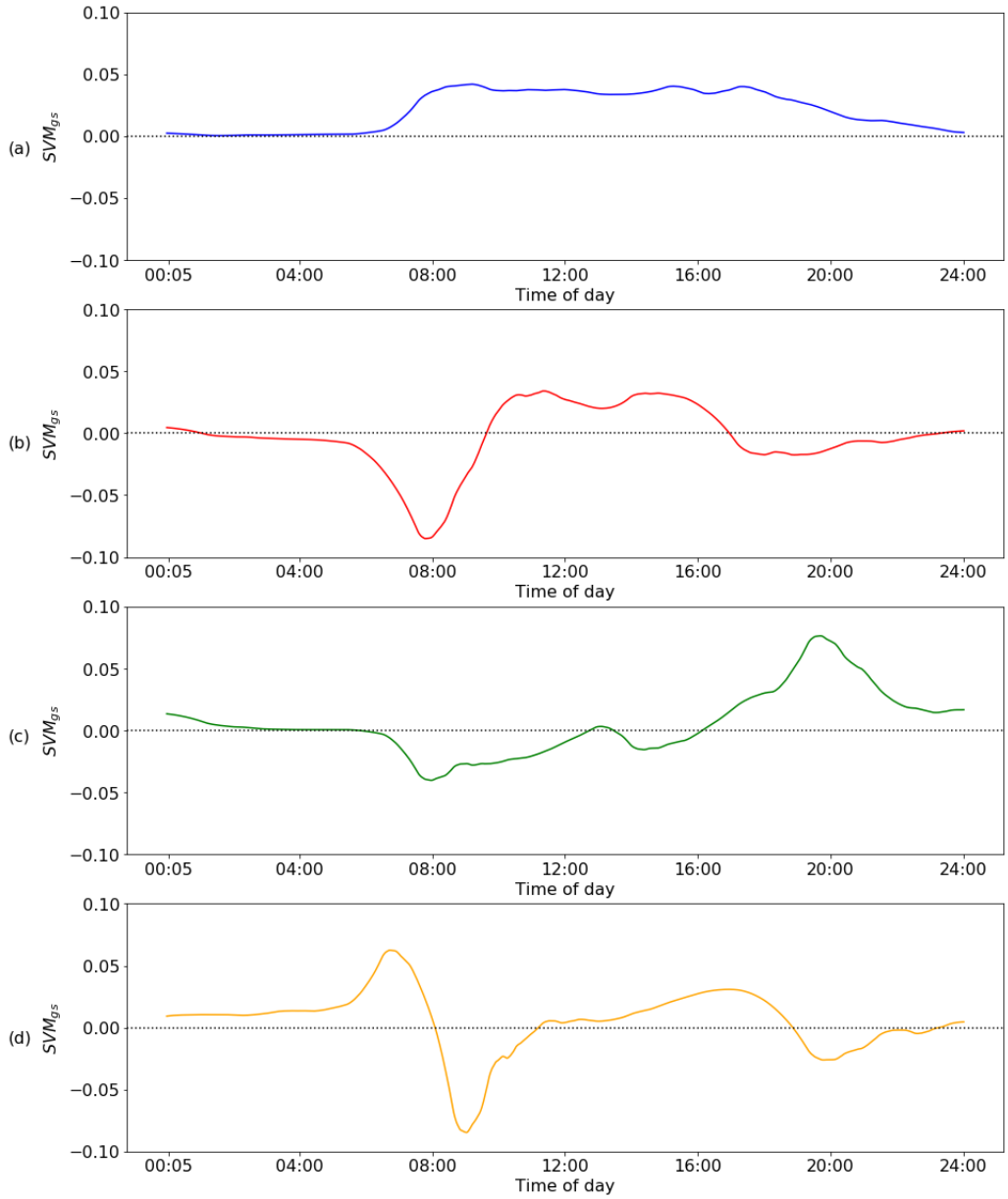


Figure 4.3: FPCs (a) ϕ_1, (b) ϕ_2, (c) ϕ_3, (d) ϕ_4

Each curve accounts for a certain percentage of variation, with the first component explaining the most variation and decreasing thereafter. The explained variance for the first 20 FPCs is given in Table 4.2.

ϕ_i	Explained Variance (%)	Cumulative Explained Variance
1	36.51	36.51
2	12.20	48.71
3	11.05	59.76
4	6.24	66.00
5	5.08	71.07
6	3.84	74.92
7	3.51	78.43
8	3.50	81.93
9	2.91	84.83
10	2.54	87.37
11	2.30	89.68
12	1.99	91.67
13	1.77	93.44
14	1.69	95.14
15	1.48	96.62
16	1.09	97.71
17	0.77	98.48
18	0.61	99.10
19	0.35	99.45
20	0.23	99.68

Table 4.2: Functional principal components explained variance

If it were standard principal components that were calculated, there would be as many principal components as there columns in our data, which in our case is 1440. As more components are added, more and more of the variation is explained. Once 1440 is reached, 100% of the variation is explained, although this would tend towards 100% much sooner than 1440. For FPCs, as seen in Table 4.2, 99% of the variance is explained with 18. To illustrate this, the cumulative explained variance is plotted in Figure 4.4.

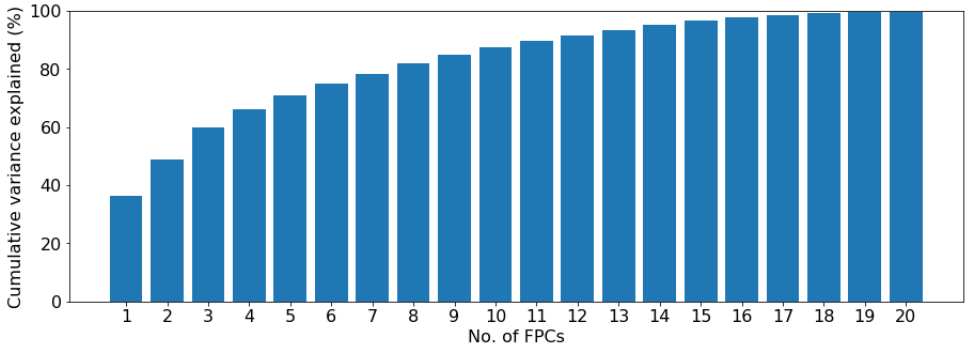


Figure 4.4: Cumulative explained variance vs. No. of FPCs

Figure 4.4 can be useful to see how far the data can be reduced, and provides a

good indication of the point of diminishing returns; the point where little variance is gained by retaining additional FPCs. A method is needed to choose the number of principal components (k) that will be used. In this study a Fraction of Variance Explained (FVE) method was used. The threshold was set such that 99% of variance is retained. Using this method, 18 FPCs were retained.

The original curve can then be reconstructed using Equation (4.3); this is illustrated for an example individual (ID 8) in Figure 4.5.

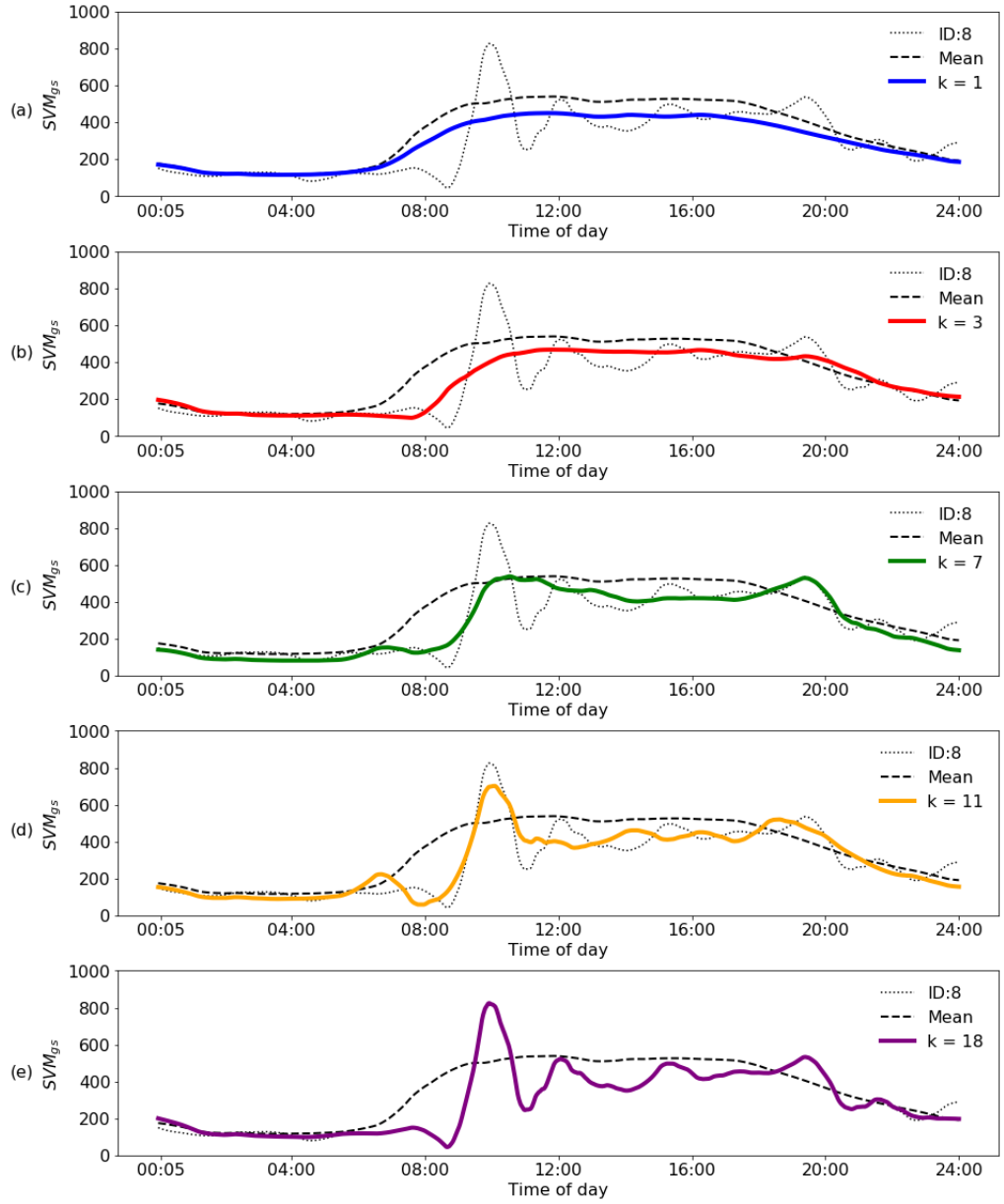


Figure 4.5: FPCA recomposition for ID 8. (a) k=1, (b) k=3, (c) k=7, (d) k=11 (e) k=18

The first FPC has a very similar shape to the mean curve, so this component is used to shift individuals up and down. ID 8 is less active than the average person in the cohort, and so the reconstruction is shifted down in Figure 4.5(a). FPCs are continually added to the reconstruction until the original curve is eventually reconstructed, shown in Figure 4.5(e). Table 4.3 gives the scores for the first 10 principal components for a sample of 10 individuals in the cohort.

ID	c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}	c_{i6}	c_{i7}	c_{i8}	c_{i9}	c_{i10}
8	-2398	1206	1462	-54	-212	-329	-1773	1953	45	-347
138	516	688	-1949	-1073	-1613	446	1334	-919	1415	-420
174	5313	488	-2885	-418	-769	1281	208	-2285	2006	-26
335	-2689	-770	-1663	1853	-1183	-1301	382	-55	102	134
719	-2094	-1468	807	-1561	-677	-63	-848	1462	377	1030
762	-212	1587	-639	-770	-937	262	-199	-322	185	309
905	-193	-1825	1613	1849	-964	-1169	-1427	-1675	1512	-1600
908	-168	1674	-677	1894	-999	617	248	18	-235	1147
914	4916	-2318	-3000	6156	1005	796	2094	385	6397	1201
1143	3742	224	-327	-2715	823	546	793	511	86	409

Table 4.3: First 10 principal component weights for first 10 IDs

To get an idea of the spread of these values, histograms for c_{ik}, k=$\{1, 2, 3, 5, 7\}$ are shown in Figure 4.6.

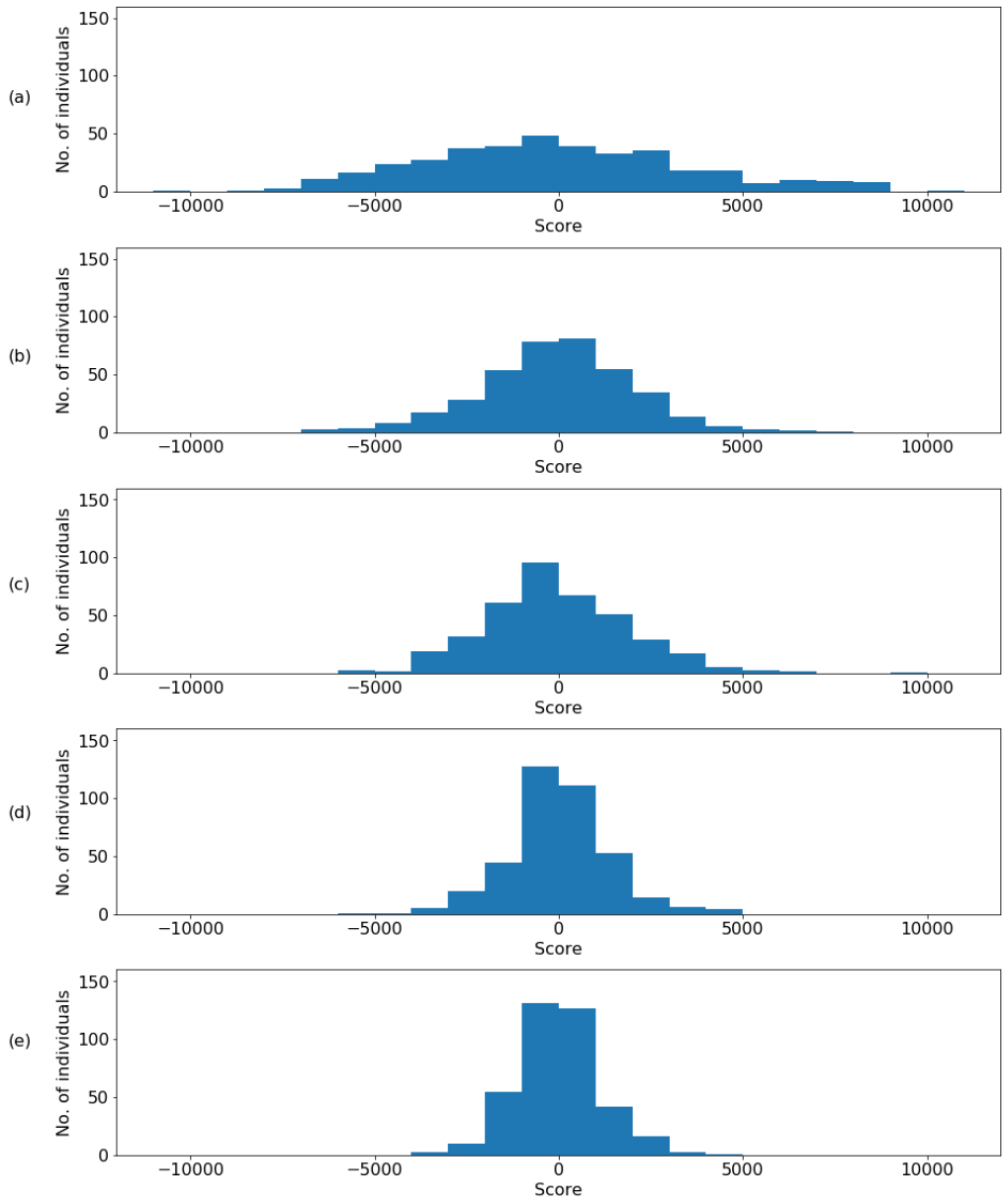


Figure 4.6: Histograms of FPC scores. (a) c_{i1}, (b) c_{i2}, (c) c_{i3}, (d) c_{i5}, (e) c_{i7}

As expected, these scores are all centred around zero, as the weight vectors were calculated to have mean zero. As we move from the first FPC (Figure 4.6(a)) to FPCs which explain less variance, this spread contracts. To make this explicit the standard deviation and absolute mean for the scores of each FPC is given in Table 4.4.

i	Standard deviation	Absolute mean
1	2918	3694
2	1634	2135
3	1547	2032
4	1133	1526
5	1027	1377
6	880	1197
7	878	1145
8	853	1143
9	740	1042
10	742	973
11	685	927
12	647	862
13	619	813
14	575	795
15	549	744
16	472	638
17	382	536
18	349	477

Table 4.4: FPC scores: Standard deviation and absolute mean

As can be seen in Table 4.4, the standard deviation and absolute mean for the weight vectors are decreasing. To further illustrate, four FPCs, +/- 1 standard deviation are shown in Figure 4.7.

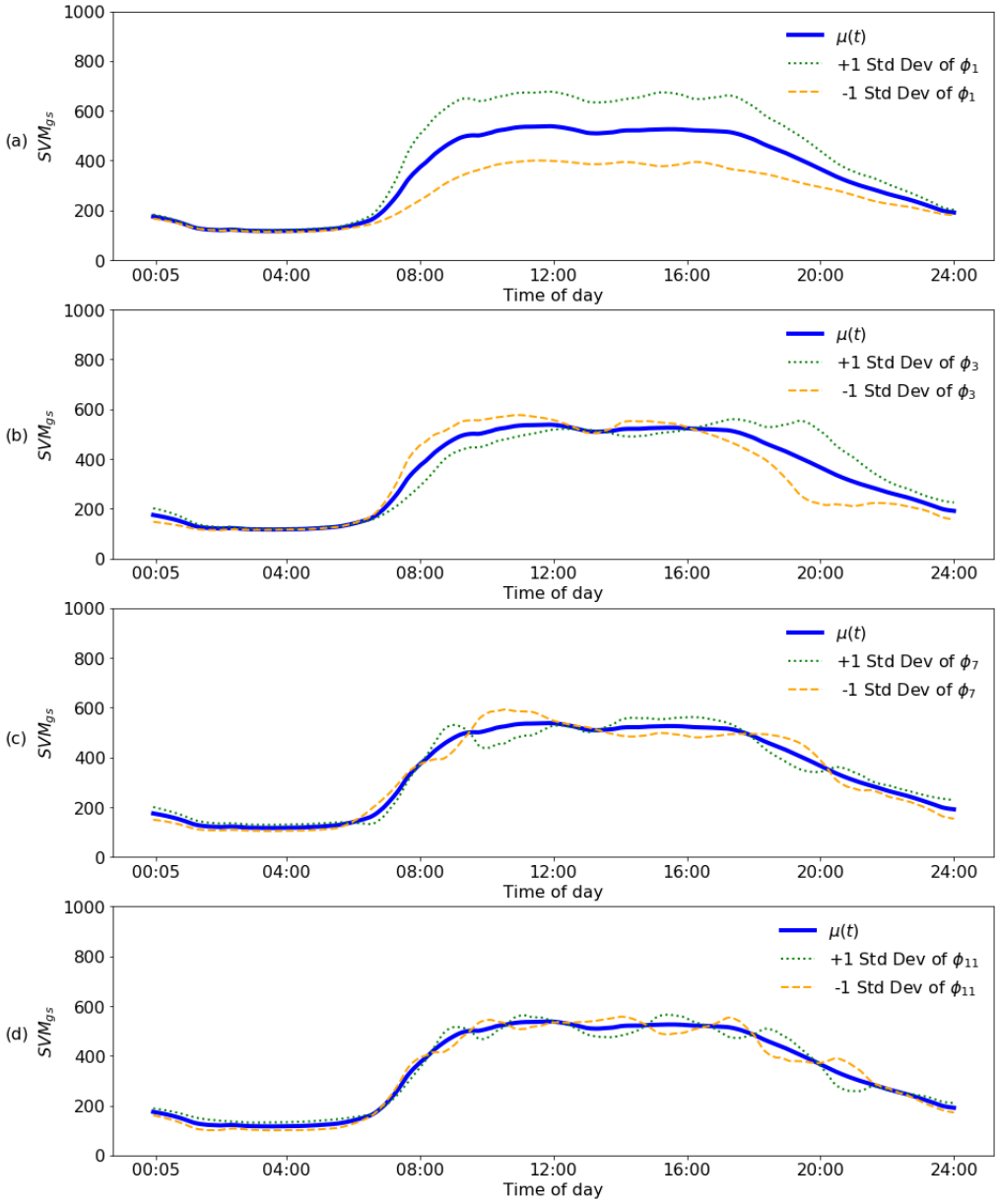


Figure 4.7: Mean curve +/- 1 standard deviation. (a) ϕ_1, (b) ϕ_3, (c) ϕ_7, (d) ϕ_{11}

In Figure 4.7(a), the curves formed by adding or subtracting one standard deviation of ϕ_1 are far removed from the mean. Whereas, in Figure 4.7(d), the curves formed by adding or subtracting one standard deviation of ϕ_{11} are very close to the mean. This demonstrates that the spread of scores has decreased.

An important aspect of FPCA is the examination of the scores, c_{ik}, of each curve on each component. The scores for the first two FPCs are plotted against each other in Figure 4.8, with the most extreme scoring individuals annotated.

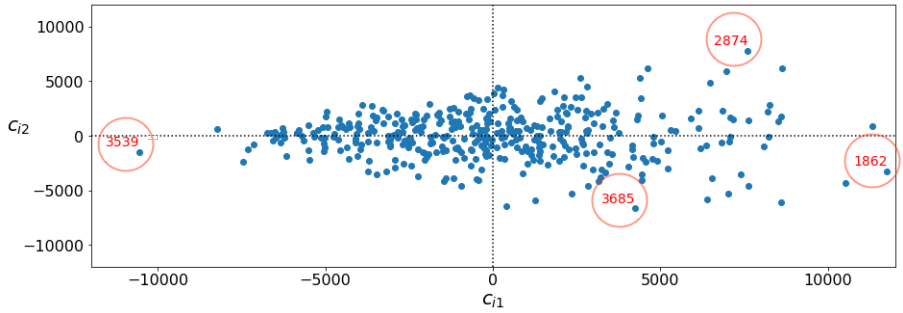


Figure 4.8: c_{i1} vs. c_{i2}

The most extreme IDs are the maximum and minimum scores on each component. ID 2874 has a large positive weighting for both the first and second FPCs. The first FPC (Figure 4.3(a)) has a similar shape to the mean, μ, (Figure 4.5(a)) and so this large positive weighting means that this particular ID's average is above the mean of the cohort. This is illustrated in Figure 4.9. In each figure, the dashed grey line represents the original curve for ID 2874. Figure 4.9(a) shows the mean curve, while Figure 4.9(b) & (c) show the effect of adding the first and second FPC respectively.

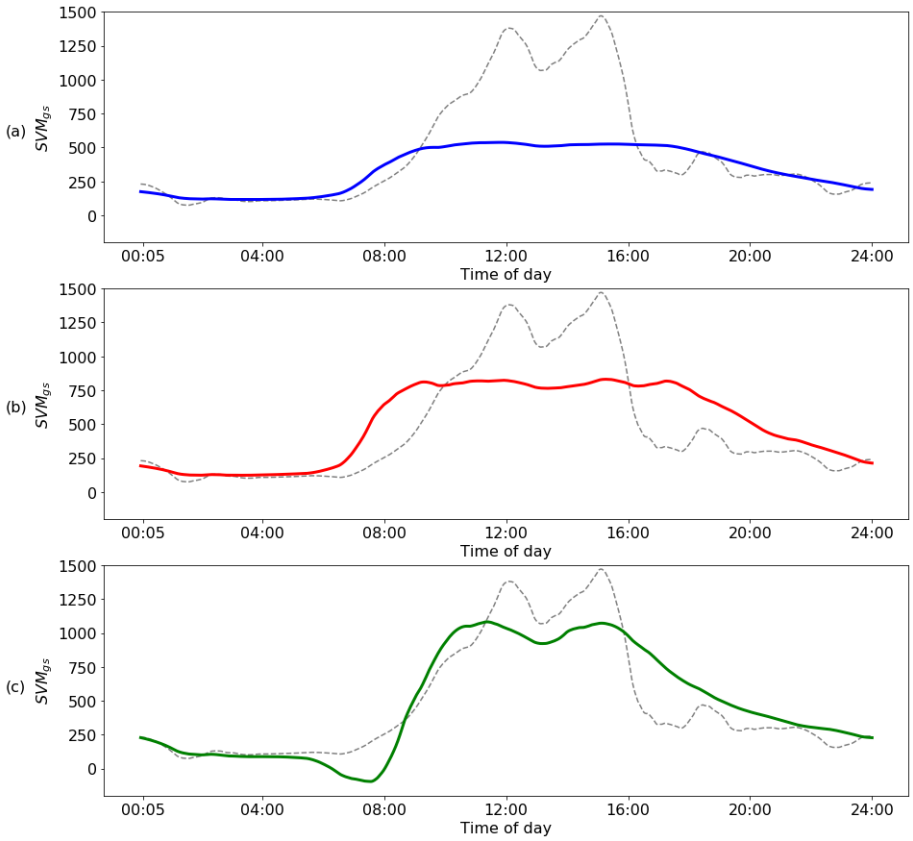


Figure 4.9: FPCA recomposition for ID 2874. (a) μ, (b) $k = 1$, (c) $k = 2$

Similarly ID 3539 has a large negative weight for the first FPC. As expected, this will shift the mean down for this individual, as shown in Figure 4.10. In each figure, the dashed grey line represents the original curve for ID 3539.

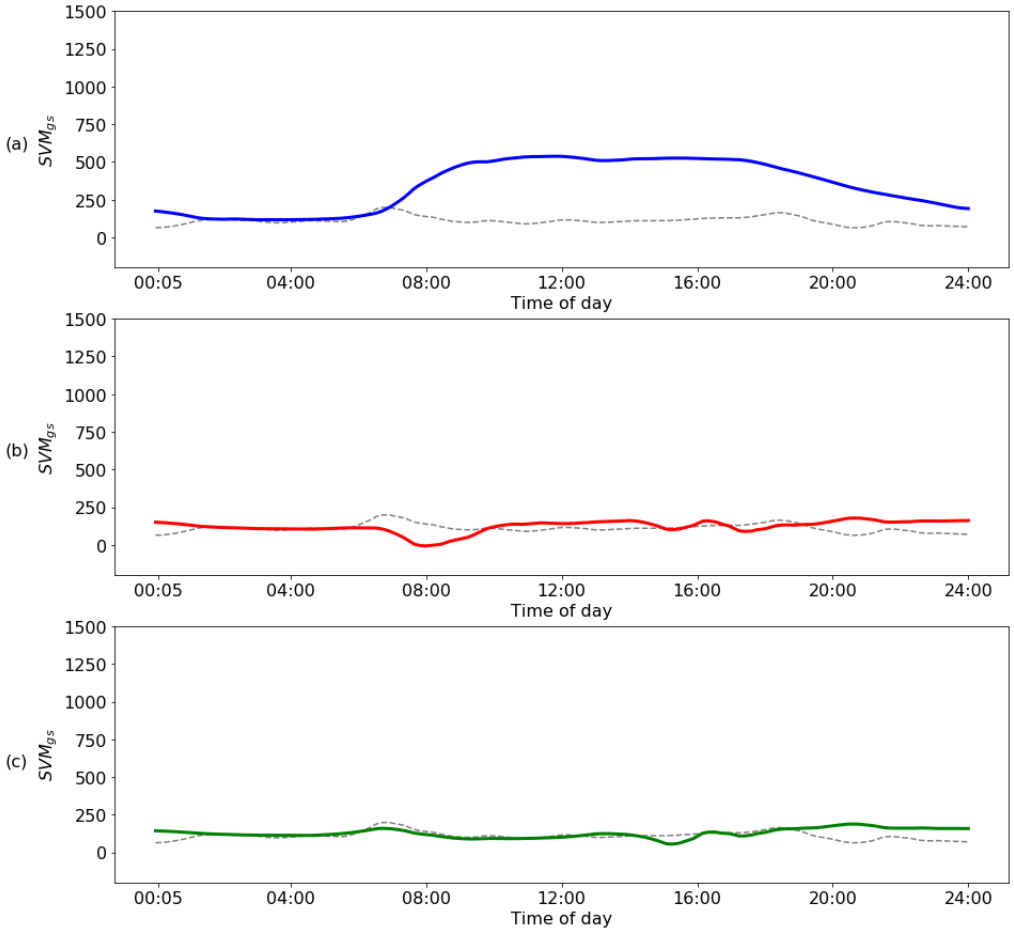


Figure 4.10: FPCA signal recomposition for ID 3539. (a) μ, (b) $k = 1$, (c) $k = 2$

These two large positive and negative scores for the first FPC (ϕ_1) seems to be indicative of low and high activity individuals. Extending this logic to the full cohort, then comparisons can be explored between these weightings and the results from the cluster analysis from Chapter 3.

In Chapter 3, five clusters were identified; namely morning larks (M-type), night owls (E-type) and 3 levels (low, moderate and high) of neither type (N-type). Considering the N-types first, it would be expected that the individuals in the high activity cluster would have large positive values for the first FPC (ϕ_1). Conversely, the constituents from the low activity group would have negative values. This is illustrated in Figure 4.11, where the coefficient weights are plotted for each individual in the cluster, note that only every 3$^{\text{rd}}$ individual is labelled in Figure 4.11(b) as there are many more individuals in that cluster as

opposed to the high activity cluster, 40 vs. 132.

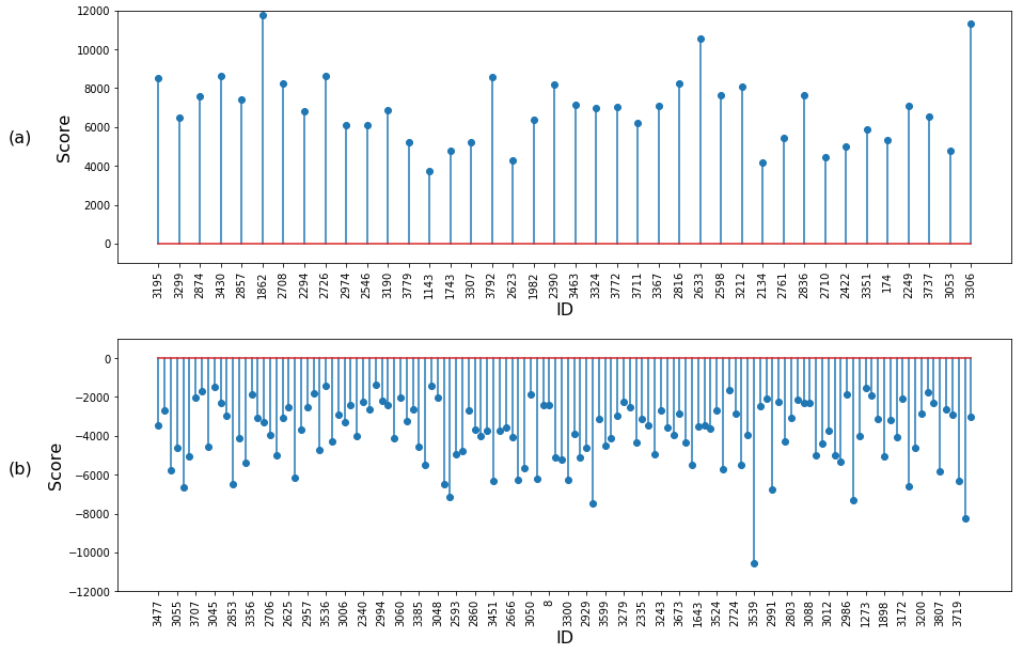


Figure 4.11: First FPC (ϕ_1) scores. (a) N-type: High, (b) N-type: Low

Given the similarities observed between these FPC scores and the results of the cluster analysis, it will be interesting to investigate the parallels between both methods. Recall from Chapter 3, that K-means clustering was based on the sum of squared euclidean distances between point observations. For clustering using the FPC scores, the pairwise distance between their scores will be used. To illustrate, the scores for ID 2874 and 3539, the individuals who scored highest and lowest on the first FPC, are plotted in Figure 4.12.

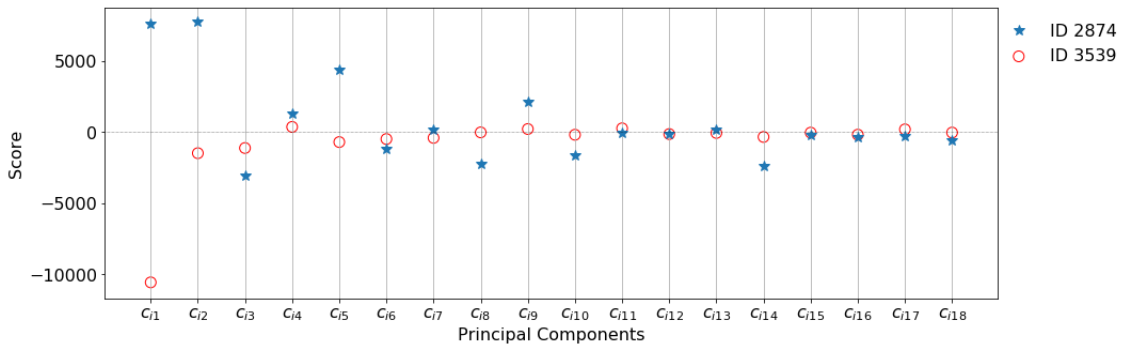


Figure 4.12: FPC scores for ID 2874 and 3539

K-means clusters individuals based on the sum of squared distances. The differences in scores will converge as higher FPCs are considered, for example

c_{i15} and c_{i18} in Figure 4.12 have negligible distance between the individuals. This was also evident in Figure 4.6, as the spread of values was seen to get smaller for the latter FPCs. For this reason, clustering will have more emphasis on the first few FPCs as their differences between individuals will be larger, hence contributing more to the sum of squared differences. Having concluded in Chapter 3, that five was the optimal number of clusters for K-means, that number will be used again here. The results from the clustering, along with the cluster sizes and the average score from the first five FPCs are given in Table 4.5.

Cluster	Size	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5
1	40	6903	68	-997	-290	-332
2	56	1813	-338	2961	-784	-15
3	64	1000	-2336	-1249	1014	251
4	96	684	1821	-471	60	66
5	132	-3844	-68	-5	-115	-62

Table 4.5: K-means on FPC scores

To visualise what these average component scores mean, they can be used to reconstruct the average profiles, similar to the centroids seen in Chapter 3. These reconstructions are shown in Figure 4.13.

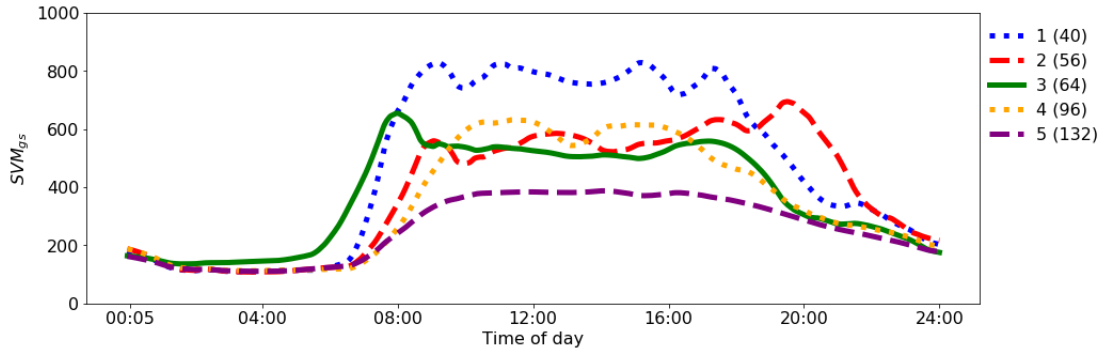


Figure 4.13: Centroid reconstruction from average FPCA scores

With the help of Figure 4.13, these clusters can again be given labels. Cluster 1 (blue) is high, cluster 2 (red) is evening, cluster 3 (green) is morning, cluster 4 (yellow) is moderate and cluster 5 (purple) is low.

To further illustrate the differences between these clusters, the average scores are plotted in Figure 4.14(a). The clusters we named high (1), moderate (4) and low (5) are highlighted in Figure 4.14(b), while the evening (2) and morning (3) clusters are shown in Figure 4.14(c) to emphasise the differences between the two.

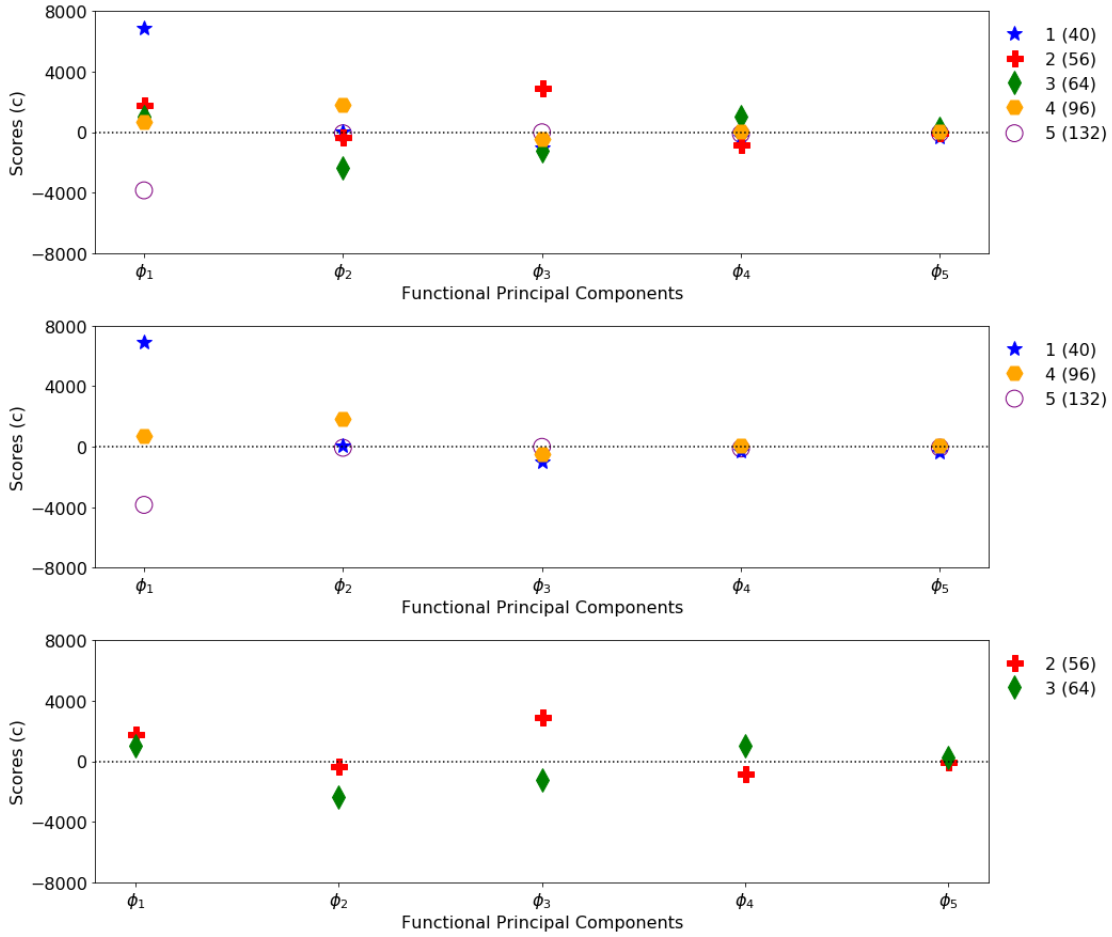


Figure 4.14: Average FPCA scores (a) All clusters, (b) High (1), Moderate (4) & Low (5), (c) Evening(2) & Morning (3)

In Figure 4.14(a), it can be seen that three clusters score similarly on the first FPC while one, the blue star, has a large positive and another, the purple circle, has a large negative. This plot just shows the averages, but obviously there is variation within each group. To highlight this, box plots are shown in Figure 4.15, again coordinated by colour, for the first FPC scores for each cluster.

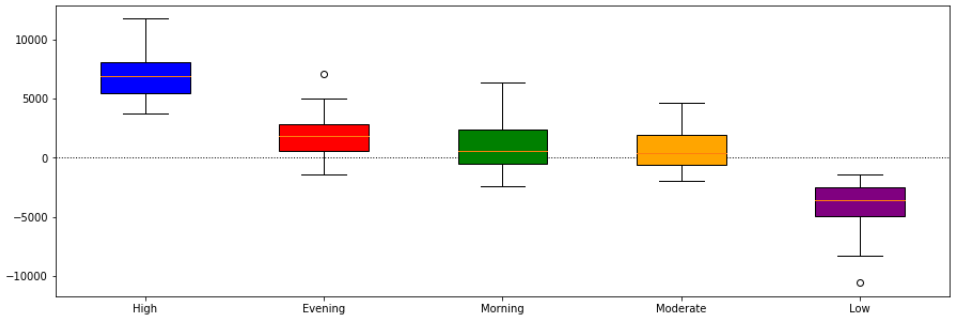


Figure 4.15: Box plots for the first FPC scores for each cluster

Every individual in the high cluster has a positive score, similarly every individual in the low cluster has a negative score, while the remaining 3 appear to be centred around zero.

To emphasize the separation between these clusters the scores on the principal components can be plotted against each other. In Figure 4.16(a) the scores on the first two components are plotted and the points are colour coordinated by cluster. The clusters are slightly mixed together in Figure 4.16(a), but if we extract the high, moderate and low clusters as shown in Figure 4.16(b), the clusters are linearly separable. This is also in the case in Figure 4.16(c), where the morning and evening clusters are extracted and the second and third FPC scores are plotted.

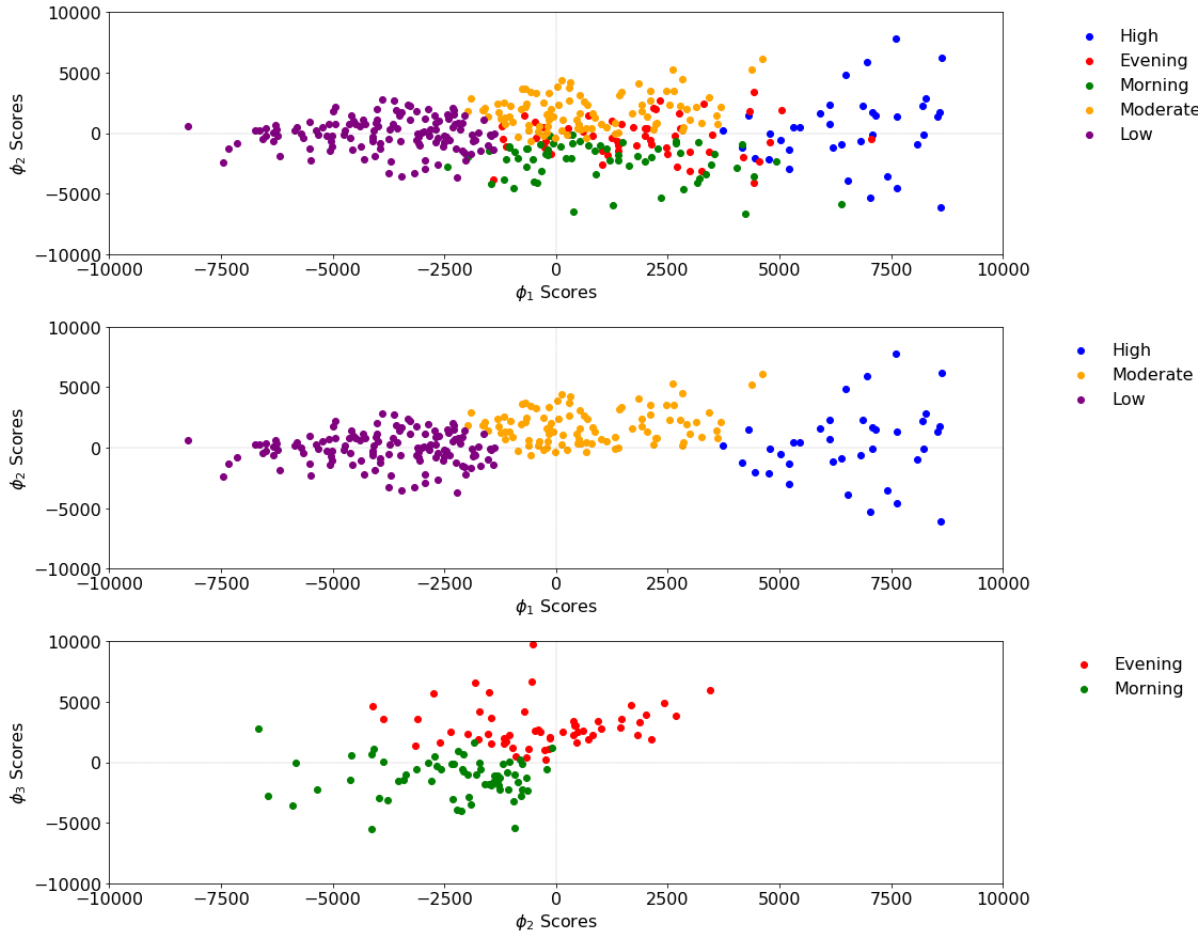


Figure 4.16: FPC comparisons (a) All clusters: ϕ_1 vs. ϕ_2, (b) High (1), Moderate (4) & Low (5): ϕ_1 vs. ϕ_2, (c) Evening(2) & Morning (3): ϕ_2 vs. ϕ_3

The method for clustering presented in this chapter differs greatly from the previous chapter. In the next section, comparison, the results from these two methods will be compared and contrasted.

4.3 Comparison

The results from the two distinct clustering methods are shown in Figure 4.17. The clusters are coordinated by both colour and number, and in addition in each cluster in the FPCA graph, Figure 4.17(b), has a unique symbol.

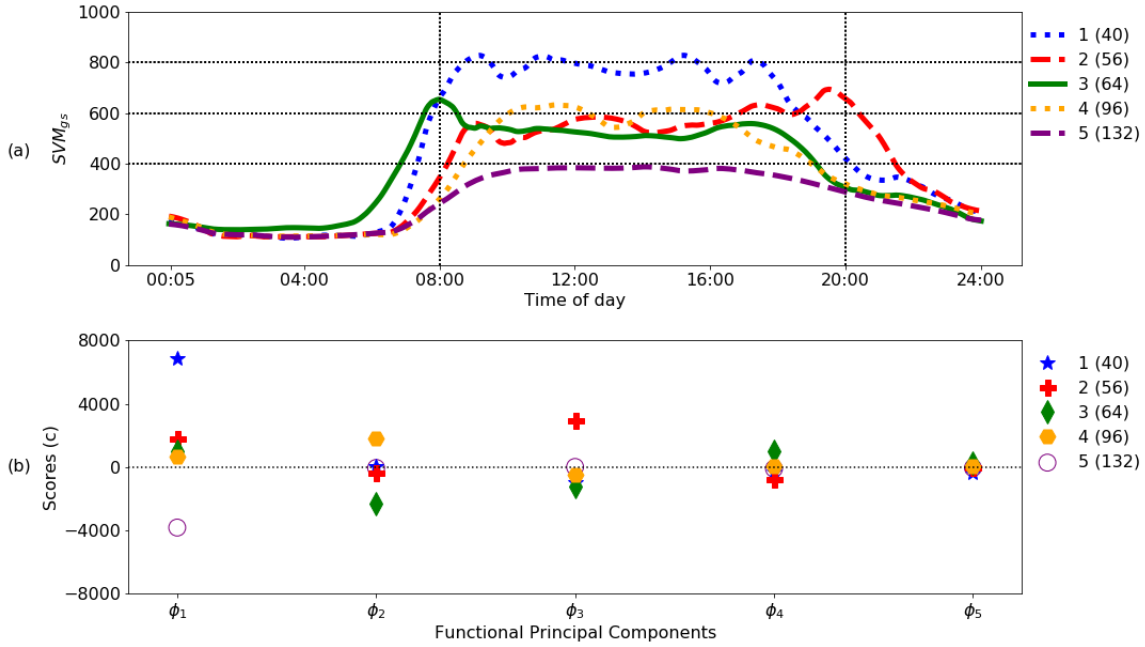


Figure 4.17: Clustering results (a) Distance, (b) FPCA

Quite surprisingly both clustering methods returned the same results. You can see in the legends of Figure 4.17 that the clusters are of the same size but the constituents still need to be verified. To picture the agreement between clustering solutions, a clustering agreement index (Aggarwal & Reddy, 2014) will be used. This quantifies the similarity between two given clusterings. The cluster agreement for the FPCA method and the distance method solutions are shown in Table 4.6.

		FPCA					
		High	Evening	Morning	Moderate	Low	Total
Distance	High	40	0	0	0	0	40
	Evening	0	56	0	0	0	56
	Morning	0	0	64	0	0	64
	Moderate	0	0	0	96	0	96
	Low	0	0	0	0	132	132
	Total	40	56	64	96	132	

Table 4.6: Cluster agreement: FPCA vs. Distance method

The values along the diagonal are the only entries populated which implies a concordance of 100% and the clustering solutions match exactly. The clusters match exactly on size, and the constituents of each cluster were verified that they too match exactly. FPCA identified features of the data and then clustering on the scores of these features yielded the same results from clustering purely on distance alone. If clustering is all that matters, then for this dataset there is no added value in using the more complicated method of FPCA over the pure distance method.

For the groups labelled high (1), moderate (4) and low (5), the parallels between both methods are simple. Distance between profiles has a direct correspondence to the scores of the first FPC. Both methods also managed to identify the evening (2) and morning (3) clusters, which warrants more investigation. This chapter addressed how scores for the second and third FPC were used to differentiate between the evening (2) and morning (3) clusters. These components were characterised by peaks or troughs in the morning or evening periods, and relatively high scores for these influenced the clustering. However, how the distance method achieved the same result was not explored. To investigate, the distances between the centroids of the high, evening and morning clusters are shown in Figure 4.18.

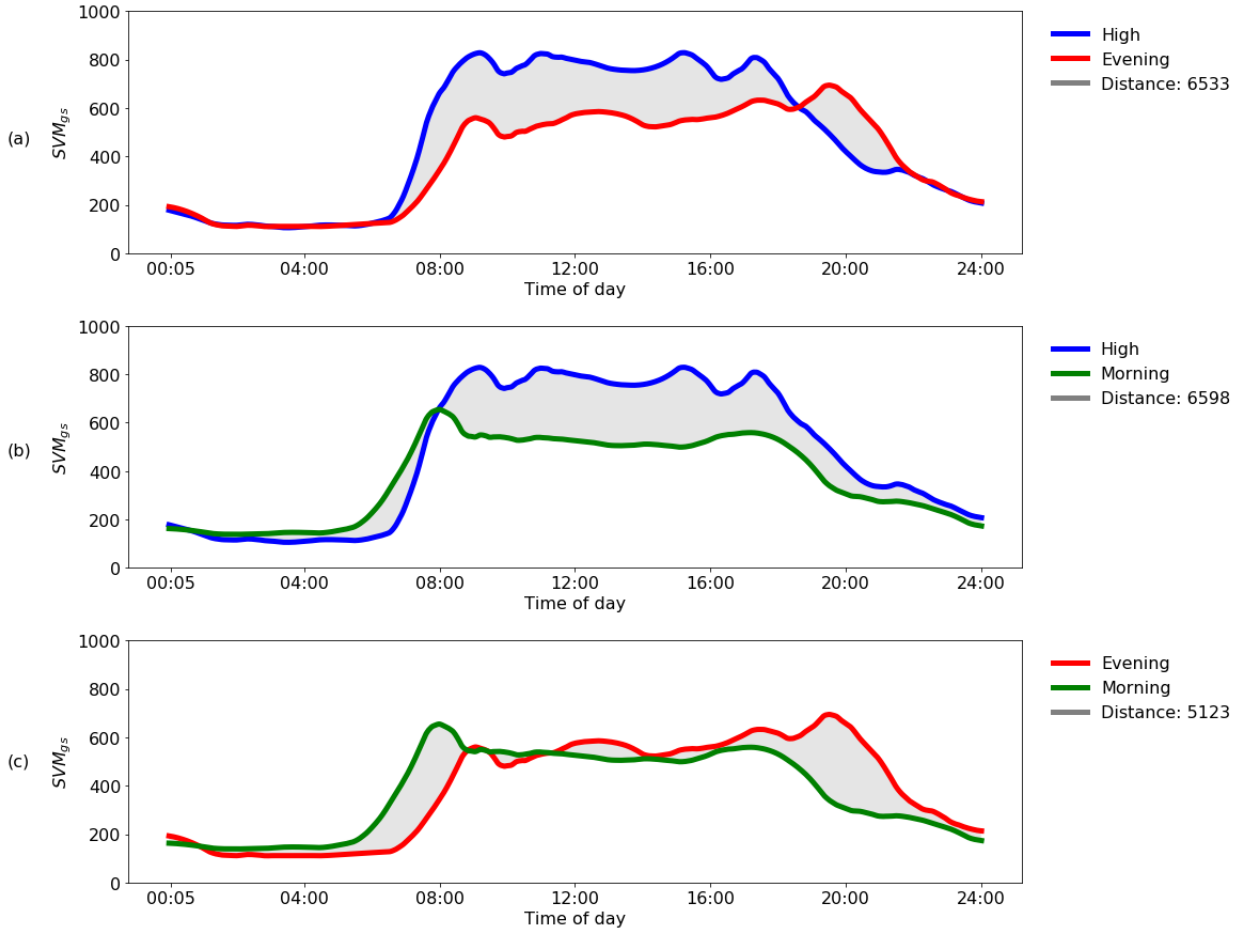


Figure 4.18: Distance between centroids (a) High (1) vs. Evening (2), (b) High (1) vs. Morning (3), (c) Evening (2) vs. Morning (3)

The evening (2) and morning (3) clusters are approximately the same distance away from the high activity cluster. For both, the majority of the contributions to the distance occur between 08:00 and 18:00. Whereas when compared to each other, in Figure 4.18(c), the largest differences occur at the tails of the day. This highlights how both the morning (3) and evening (2) clusters are separated from the high (1) activity, and how they are separated from each other. The logic for these separation can be extended for the moderate (4) and low (5) activity clusters.

FPCA decomposes the variation between individual curves into uncorrelated, temporal features. The usefulness of this depends on how these components are interpreted. When looking at an outcome it is easier to observe and interpret scores on specific principal components rather than checking to see what cluster an individual belongs to. For example, if there was a targeted physical activity intervention aimed at those who are more active in the morning, the scores for the related components would only need to be examined. FPCA also makes use

of the underlying patterns in the data which is a more intuitive way of describing these profiles.

For these reasons the FPCA will be chosen over the simple distance method. The output from the clustering based on FPCA will be used as the basis for our sensitivity analysis in Chapter 5.

Chapter 5 - Sensitivity Analysis

Sensitivity analysis is defined as the study of how the uncertainty in the output of a model (numerical or otherwise) can be apportioned to different sources of uncertainty in the model input (Saltelli, Tarantola, Campolongo, & Ratto, 2004). It is the process of recalculating outcomes under alternative assumptions to determine the impact of the input variable. Put simply, it involves changing the model and observing the resultant behaviour.

By performing such an analysis the robustness of the results of a model can be tested. It will also further the understanding of the relationships between the inputs and outputs. By doing this, it will help us observe how sensitive the results are to modelling assumptions. This will not be a comprehensive sensitivity analysis, but rather it is driven by the topics discussed in the preceding chapters. This will include varying the number of FPCs used to perform the clustering, exploring different smoothing techniques, consideration of different epoch lengths and finally the choice to use only the weekday data in our aggregation.

The results of the clustering, which utilised FPCA, will be used to gauge the sensitivity of each choice. Therefore the number of FPCs used will be the first choice to be investigated. To obtain the results, 18 components were used, but could the same clusters be identified with fewer components? The degree of overlap between our original solution and those generated using fewer FPCs will then be assessed. In other words, the number of individuals who remain in the same cluster and those who move will be illustrated.

Next was the choice of smoothing technique. The 6^{th} scale approximation from a DWT that used a DB-4 mother wavelet was used. In this sensitivity analysis, another choice of wavelet, DB-8, will be explored, along with the rougher 4^{th} and 5^{th} scale approximations. Splines, with both 11 and 23 uniform knots, will also be explored to see if their use would have affected the results. The choice of smoothing technique may then alter the FPCs that were derived, and subsequently affect the related clustering results.

The decision to use the 6^{th} scale approximation, was a direct consequence of the number of data points in a profile. Collapsing to 1 minute epochs meant that there were 1440 data points that needed to smoothed to reveal the underlying functional nature of the data. Using 5 minute epochs, meaning 288 data points, would require the use of a different scale approximation. This choice of epoch length will be explored to see if there is any information lost by using the more aggregated 5 minute epoch.

The final input to be analysed will be the number of days chosen. In our

analysis, having collapsed the data into 1 minute epochs, each epoch was then averaged over 5 days, Monday to Friday. However the participants in this study wore the accelerometer for a full week. In Chapter 1, justification was given for the choice to use only the week day data, as people's week day activity profiles often differ from those observed at the weekend. It is important to use all available data, so the inclusion of the weekend data to create the average activity profiles will also be considered.

5.1 FPCA

This section will focus on whether clustering on fewer principal components will make a difference. The scores for the first 18 components were used to perform the clustering in the previous chapter as they explained 99% of the variance. However when characterising the clusters that were formed, the scores for the first 3 components were used to describe the differences. The first principal component was used to shift the mean up and down, and was identified as the main differentiator between the low, moderate and high activity groups. While the second and third components helped distinguish between the morning and evening groups. For this reason, clustering with just the scores from the first three components will firstly be compared to the full solution. The reconstructed centroids are shown in Figure 5.1

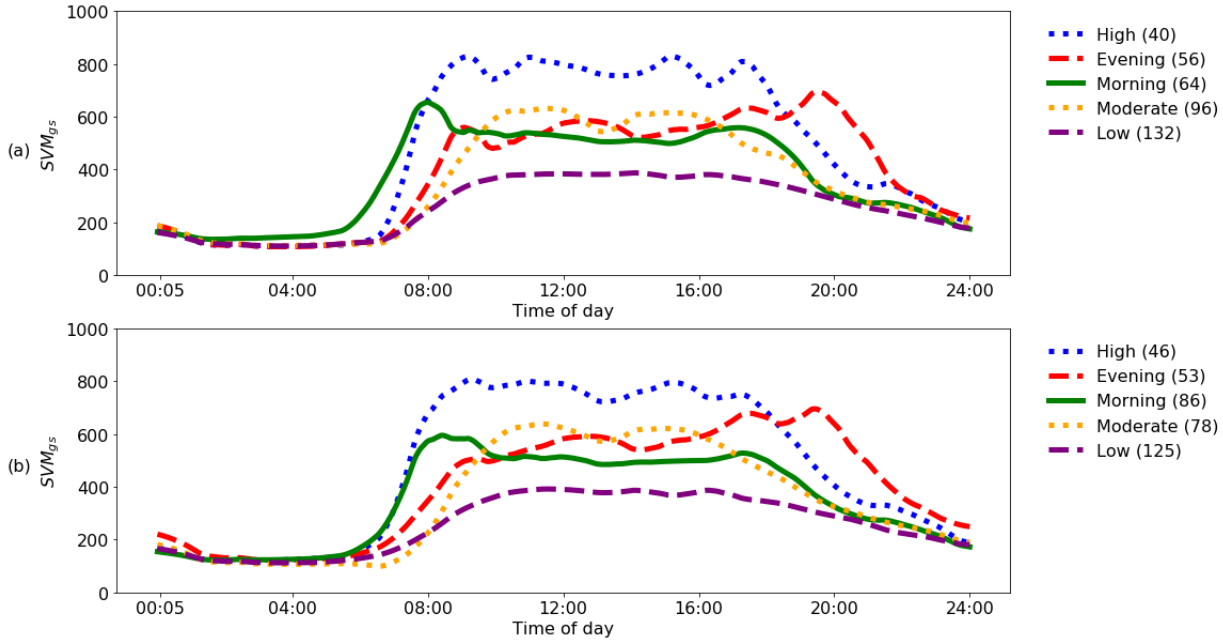


Figure 5.1: Cluster centroids (a) 18 components, (b) 3 components

The results from clustering on three components appear to give a reasonable estimate, but lacks the detail from the original result. The loss of information inherent in using a reduced number of components will be investigated. The size of

the clusters are shown in the legend in Figure 5.1. For example, the high activity cluster, has 40 constituents when 18 components were used, and 46 constituents with three components. To picture the agreement between clustering solutions, a clustering agreement index can be used again, as seen in Chapter 4. This quantifies the similarity between two given clusterings. The cluster agreement for the 18 vs. 3 components solution is shown in Table 5.1.

		18 FPCs					
		High	Evening	Morning	Moderate	Low	Total
3 FPCs	High	39	0	5	2	0	46
	Evening	1	48	2	2	0	53
	Morning	0	8	57	14	7	86
	Moderate	0	0	0	78	0	78
	Low	0	0	0	0	125	125
	Total	40	56	64	96	132	

Table 5.1: Cluster agreement: 18 vs. 3 FPCs

Of the 40 people that the 18 FPCs solution classified as high activity, the 3 FPCs solution classified 39 people the same way, with the remaining person classified as evening. The diagonal of Table 5.1 shows the number of individuals classified the same by both solutions. Dividing the sum of this diagonal by the total number of individuals, gives a concordance of 89% between the two approaches.

The inclusion of one or more of the components higher then the 3rd causes the differences in the clustering solutions. To investigate where these differences occur, the number of components will be decreased incrementally and compared with the 18 FPCs solution. To begin, the difference between 17 vs. 18 FPCs is explored, this is shown in Table 5.2.

		18 FPCs					
		High	Evening	Morning	Moderate	Low	Total
17 FPCs	High	40	0	0	0	0	40
	Evening	0	56	0	0	0	56
	Morning	0	0	64	0	0	64
	Moderate	0	0	0	96	0	96
	Low	0	0	0	0	132	132
	Total	40	56	64	96	132	

Table 5.2: Cluster agreement: 18 vs. 17 FPCs

The values along the diagonal are the only entries populated which implies a concordance of 100%. The concordance measures for every number of FPCs is shown in Table 5.3.

FPCs	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3
%	100	100	100	100	99	99	99	99	99	99	99	98	98	97	89

Table 5.3: Concordance percentages: 17 to 3 FPCs

The first reduced number of components where differences are observed is 13. What this means, is that the inclusion of the 14$^{\text{th}}$ principal component had an influence on the clustering. Again the cluster agreement index can be created to observe where the differences occur, this is shown in Table 5.4.

		18 FPCs					
		High	Evening	Morning	Moderate	Low	Total
13 FPCs	High	40	0	1	1	0	42
	Evening	0	56	0	0	0	56
	Morning	0	0	63	0	0	63
	Moderate	0	0	0	95	0	95
	Low	0	0	0	0	132	132
	Total	40	56	64	96	132	

Table 5.4: Cluster agreement: 18 vs. 13 FPCs

Encompassing the 14$^{\text{th}}$ FPC in the clustering resulted in two individuals being clustered differently. One moved from the morning cluster to the high, and the other from moderate to high. To investigate further, the IDs for these two individuals will be required, these are given in Table 5.5.

ID	18 FPCs	13 FPCs
2356	Morning	High
2676	Moderate	High

Table 5.5: Cluster change IDs: 18 vs 13 FPCs

The 14$^{\text{th}}$ FPC is shown in Figure 5.2.

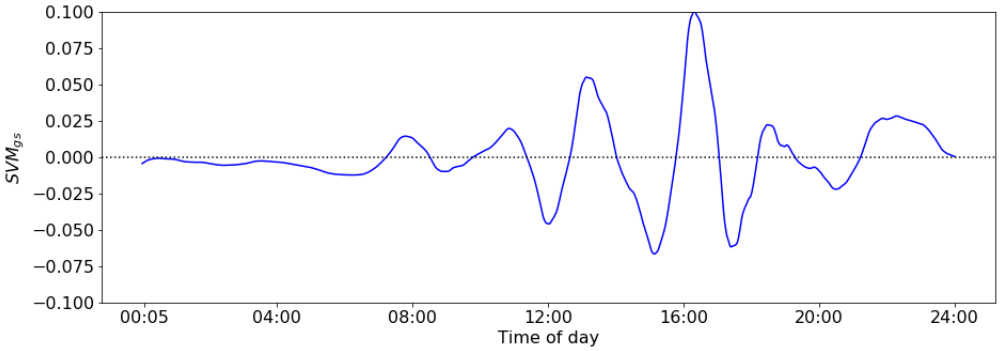


Figure 5.2: 14$^{\text{th}}$ FPC

The respective scores for this FPC can be checked for these individuals and the original curve can be reconstructed with 13 and 14 FPCs to see the effect of including this FPC. This is shown for the 2 individuals who changed clusters in Figure 5.3.

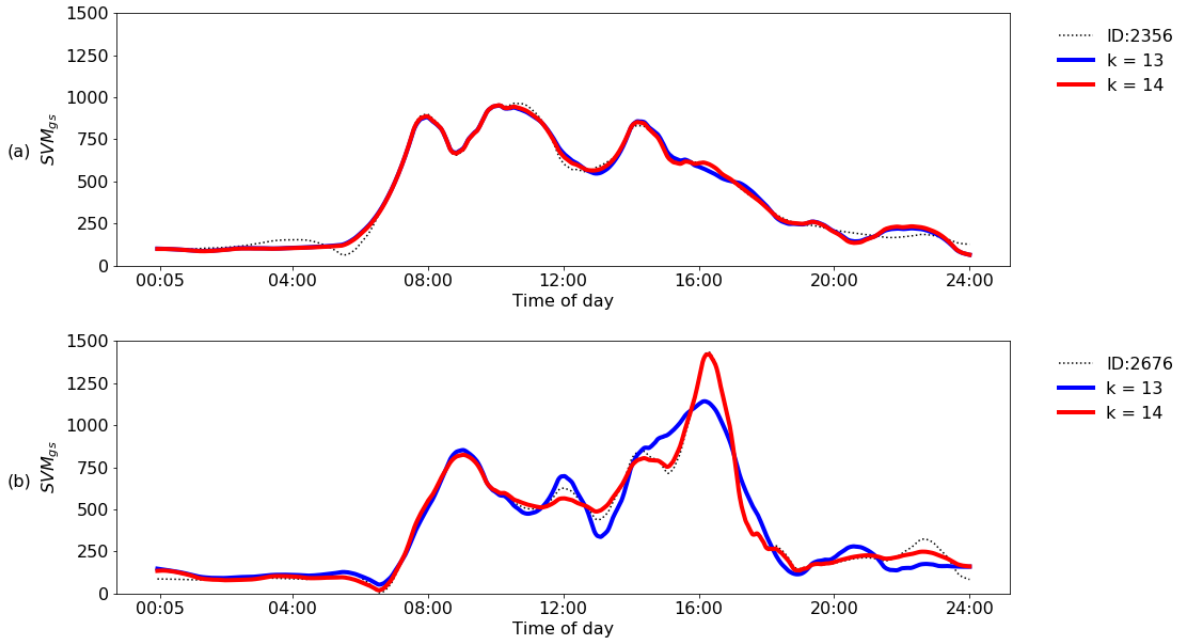


Figure 5.3: Reconstruction with 13 and 14 FPCs. (a) 2356, (b) 2676

The 14^{th} FPC has a clear effect on the profile for ID 2676, Figure 5.3(b), and the increased amplitude at around 4p.m. is conceivably enough to move the individual from the moderate to high cluster. The effect on ID 2356, Figure 5.3(a), is less clear however. In Chapter 3, the silhouette analysis demonstrated that the clusters were not well separated, so ID 2356 could potentially be a border line case and the seemingly marginal effect of adding the 14^{th} FPC could be sufficient for its clustering to change. To investigate further, the 13 and 14 FPCs reconstructions are plotted again, this time with the cluster centroids for both the morning and high clusters. This is shown in Figure 5.4.

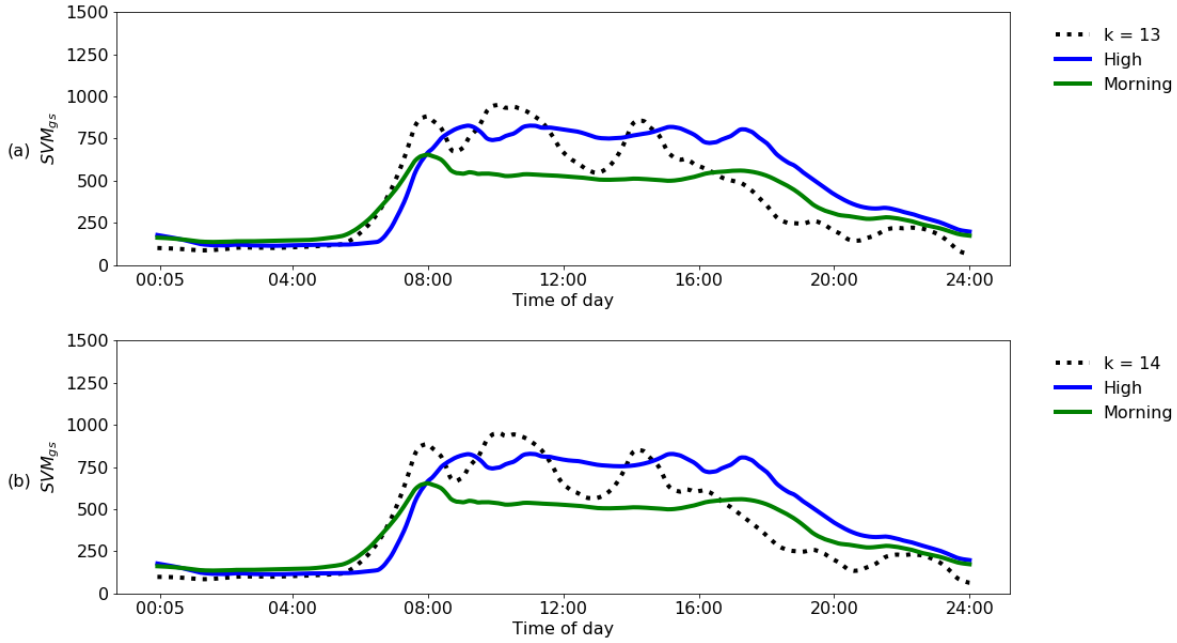


Figure 5.4: Curve reconstruction for ID 2356. (a) 13 FPCs, (b) 14 FPCs

The argument could be made for this individual belonging to either cluster. The profile has the increase in activity at around 6a.m. which is characteristic of the morning cluster, while its activity during the day is high enough that it could feasibly belong to the high activity cluster.

In Table 5.3, for every choice of the number of FPCs used (except for 3 FPCs), the concordance percentage remains close to 100%. This demonstrates that the clusters are indeed robust to the choice of the number of FPCs used. In order to run FPCA as opposed to regular PCA, the individuals needed to be described by functions or curves rather than a discrete time series of observations. To do this, smoothing techniques were applied so the next section will focus on the robustness of the clustering to the choice of smoothing method.

5.2 Smoothing

In Chapter 2, the discrete wavelet transform was chosen as the smoothing technique. Specifically the level 6 approximations of a DB-4 mother wavelet. At the end of that chapter, other methods were noted as offering similar representations of the data. So this section will explore the sensitivity of the clustering to the choice of this smoothing technique. Changing the smoothing techniques may mean that the curves used to describe individuals will be different, and hence may result in different FPCs. Therefore FPCA will be rerun for each smoothing technique, and 18 FPCs will again be used to perform the clustering.

Firstly cubic splines with both 23 and 11 uniform knots (every 1 hour and every 2 hours respectively) will be implemented and analysed. A change of basis using our principal components will be performed, to transform these into subject specific scores which will then be clustered. These results will be compared to the previous results to determine the sensitivity of this choice of input.

Then a similar wavelet, the DB-8 mother, will be contrasted. It was shown in Chapter 2 how these two wavelets differ when fitting a curve for a specific individual, so how they affect the clustering as a whole will be explored in this chapter. DB-4 has fewer vanishing moments, and therefore more compact support, when compared to DB-8.

The first step is to fit every individual's data with a cubic spline, with both 23 and 11 uniform knots. Then FPCA is run to generate the subject specific scores. These scores are then used to perform the clustering with K-means, with K=5. With that complete, the cluster centroids can be plotted, this is shown in Figure 5.5. Again the number of constituents in each cluster is shown in the legend.

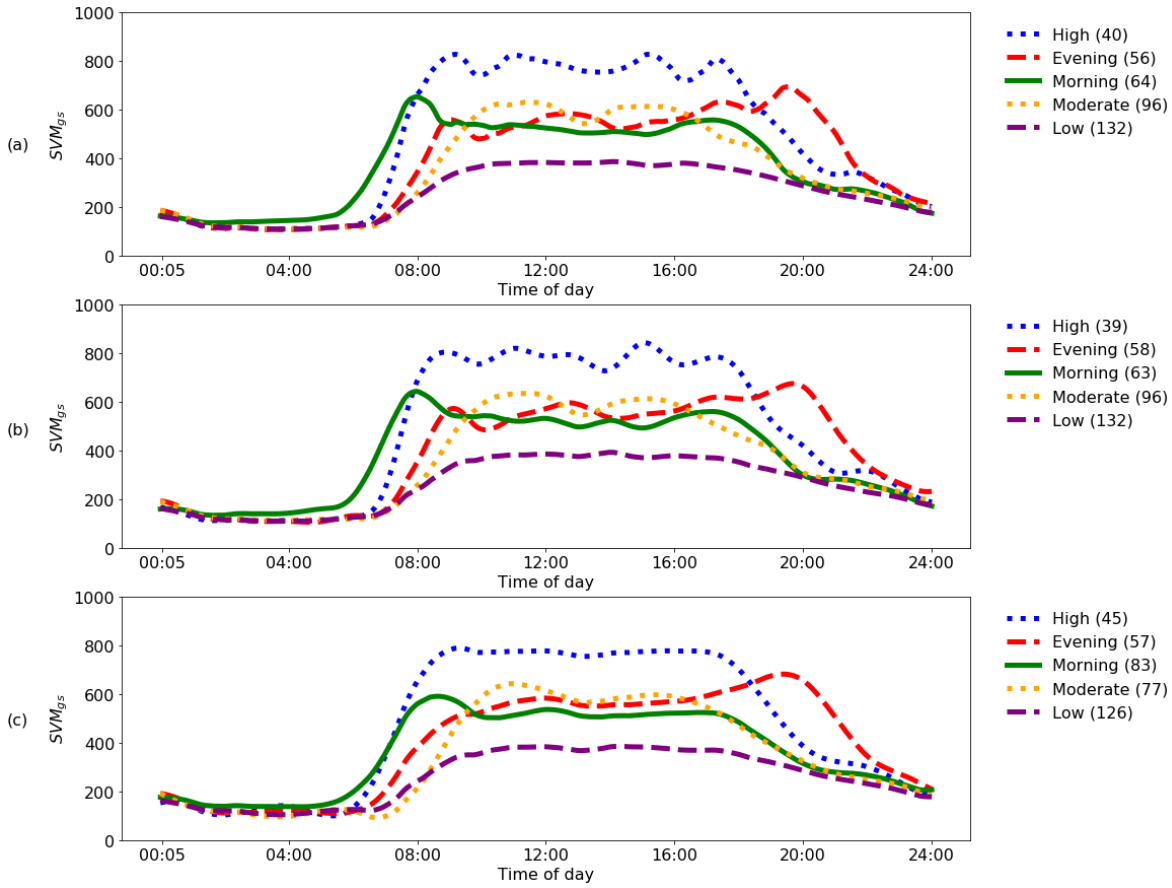


Figure 5.5: Cluster centroids (a) DB-4, (b) Cubic spline - 23 knots, (c) Cubic spline - 11 knots

Judging solely by the shapes of the centroids and the size of each cluster, the cubic splines with 23 knots appears to have performed similarly to our chosen wavelet method. The cubic splines with 11 knots have flatter and smoother centroids, and they are not picking up the nuances that the other two methods do. This was to be expected as 11 knots means that there is a much wider support, and so some of the details are lost. The differences in clustering after cubic splines with 23 knots (CS23) can be investigated further using a cluster agreement index. This is shown in Table 5.6.

		DWT - DB4					
		High	Evening	Morning	Moderate	Low	Total
CS23	High	38	0	0	1	0	39
	Evening	2	56	0	0	0	58
	Morning	0	0	63	0	0	63
	Moderate	0	0	1	95	0	96
	Low	0	0	0	0	132	132
	Total	40	56	64	96	132	

Table 5.6: Cluster agreement: DWT - DB4 vs. Cubic spline (23 knots)

This shows that 4 individuals changed cluster. Their IDs are shown in Table 5.7.

ID	DWT- DB4	CS23
1143	High	Evening
2294	High	Evening
2676	Moderate	High
3697	Morning	Moderate

Table 5.7: Cluster change IDs: DWT - DB4 vs. Cubic spline (23 knots)

ID 2676 again moves from the moderate to the high cluster, as it did in section 5.1. Two individuals move from high to evening, and one moves from the morning to the moderate cluster. It was observed in Figure 5.5, that different smoothing techniques will result in different cluster centroids. Therefore to explore the differences in the clustering solutions, the fitted curves for these individuals will need to be compared to their respective cluster centroids for each method. The two individuals that went from high to evening will be explored. The fitted lines using both smoothing methods, along with the centroids for the high and evening clusters are displayed in Figure 5.6.

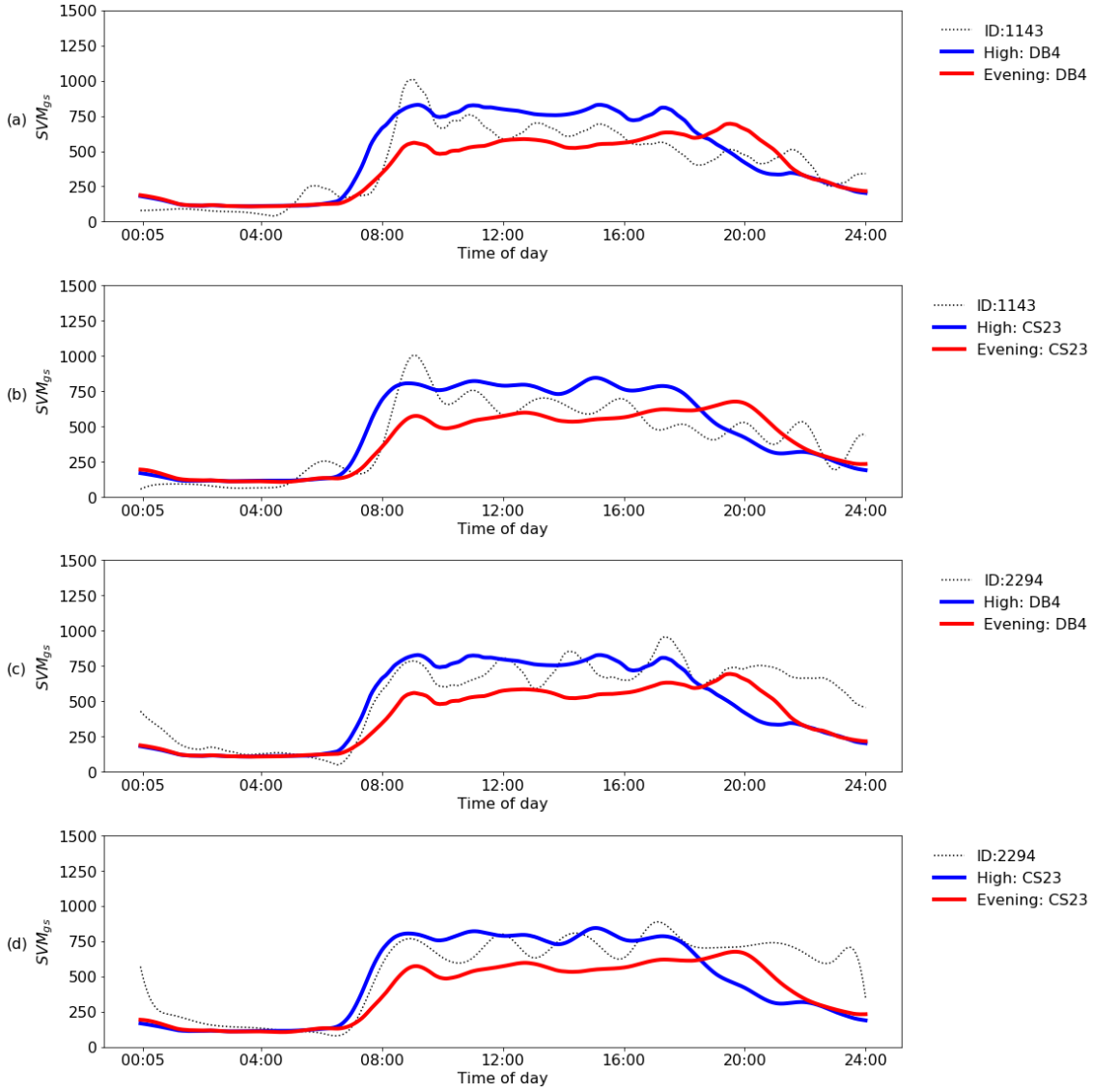


Figure 5.6: Cluster constituent changes (a) ID:1143, DB4, (b) ID:1143, CS23, (c) ID:2294, DB4, (d) ID:2294, CS23

The evening cluster is characterised by increased activity at around 8p.m., and it is around this point where differences in the profiles are observed. In Figure 5.6(b), where the cubic spline method is used, there is a slight peak at this time, which is not as pronounced when the wavelets are used, Figure 5.6(a). Conversely, in Figure 5.6(c), there is a slight trough prior to 8p.m., which is not present in Figure 5.6(d), where the cubic splines are used. These profile attributes are enough to affect the clustering of these individuals. Only 4 individuals changing cluster still yields a concordance of 99%, so overall the cluster solution is mostly unaffected by the difference between these smoothing techniques. Next a different choice of wavelet will be looked at.

Using a DB-8 mother wavelet in the DWT yields the same clusters. This is displayed in a cluster agreement index in Table 5.8.

		DB4					
		High	Evening	Morning	Moderate	Low	Total
DB8	High	40	0	0	0	0	40
	Evening	0	56	0	0	0	56
	Morning	0	0	64	0	0	64
	Moderate	0	0	0	96	0	96
	Low	0	0	0	0	132	132
	Total	40	56	64	96	132	

Table 5.8: Cluster agreement: DWT - DB4 vs. DWT - DB8

The DWT decomposes a curve into a series of approximations and details, and these detail coefficients were set to zero in our implementation. The choice was made to use the 6^{th} scale approximations as it provided a smooth representation of the underlying functional nature of the data. To illustrate the effect of using the 4^{th} and 5^{th} scale approximations, they are plotted along with the 6^{th} scale for an example individual in Figure 5.7.

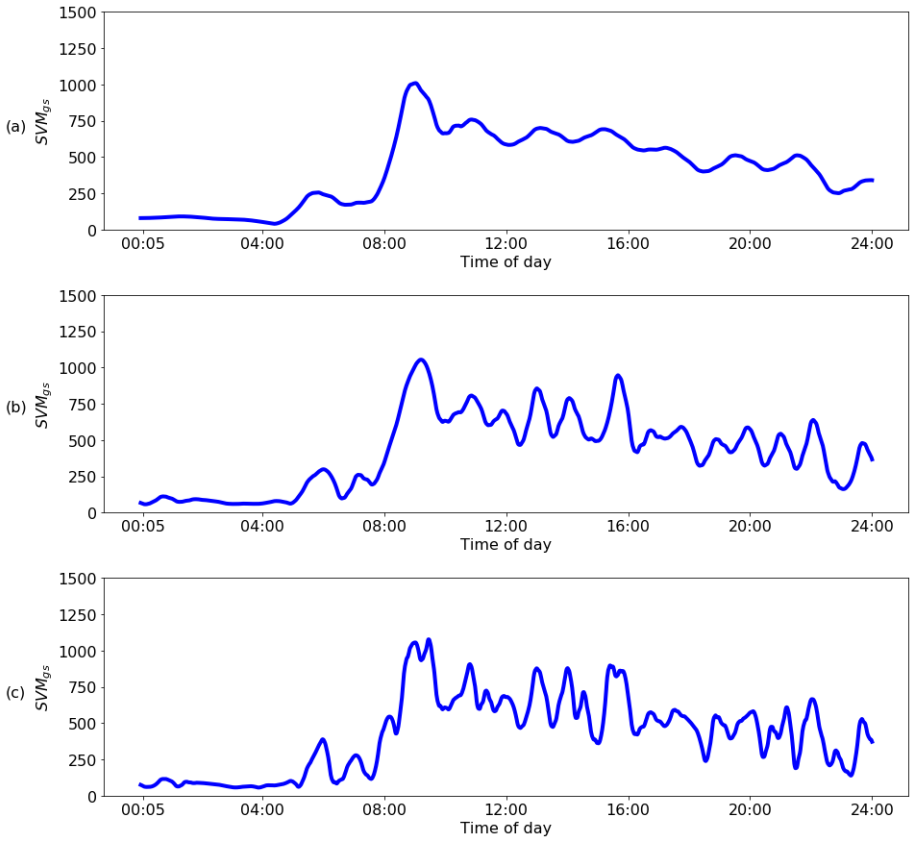


Figure 5.7: ID:1143 scale approximations. (a) 6^{th}, (b) 5^{th}, (c) 4^{th}

It is clear that the representations are less smooth, but how would their use affect the clustering? Once more, each person's profile/curve has changed so FPCA will need to be rerun and Kmeans implemented on the subject specific scores. The concordance measures for each are shown in Table 5.9.

Scale approximation	5^{th}	4^{th}
%	99	98

Table 5.9: Concordance percentages: 5^{th} and 4^{th} scale approximations

Every choice of smoothing technique has demonstrated a relatively high concordance and so it can be concluded that the clusters are robust to this choice. The different scale approximations represent alternative ways of aggregating the data, and it has been shown that the clusters are robust to this choice. In the preprocessing of the data, the raw data was collapsed into 1 minute epochs before the activity profiles were created. Section 5.3 will explore an alternative epoch length, 5 minutes, to see if there is any loss of information.

5.3 Epoch

Using a epoch length of 1 minute, resulted in a time series of length 1440 for every individual. If a 5 minute epoch is used, it means the time series will have length 288. Therefore, a different scale approximation needs to be used. The 3rd, 4th and 5th scale approximations for an example individual are shown in Figure 5.8.

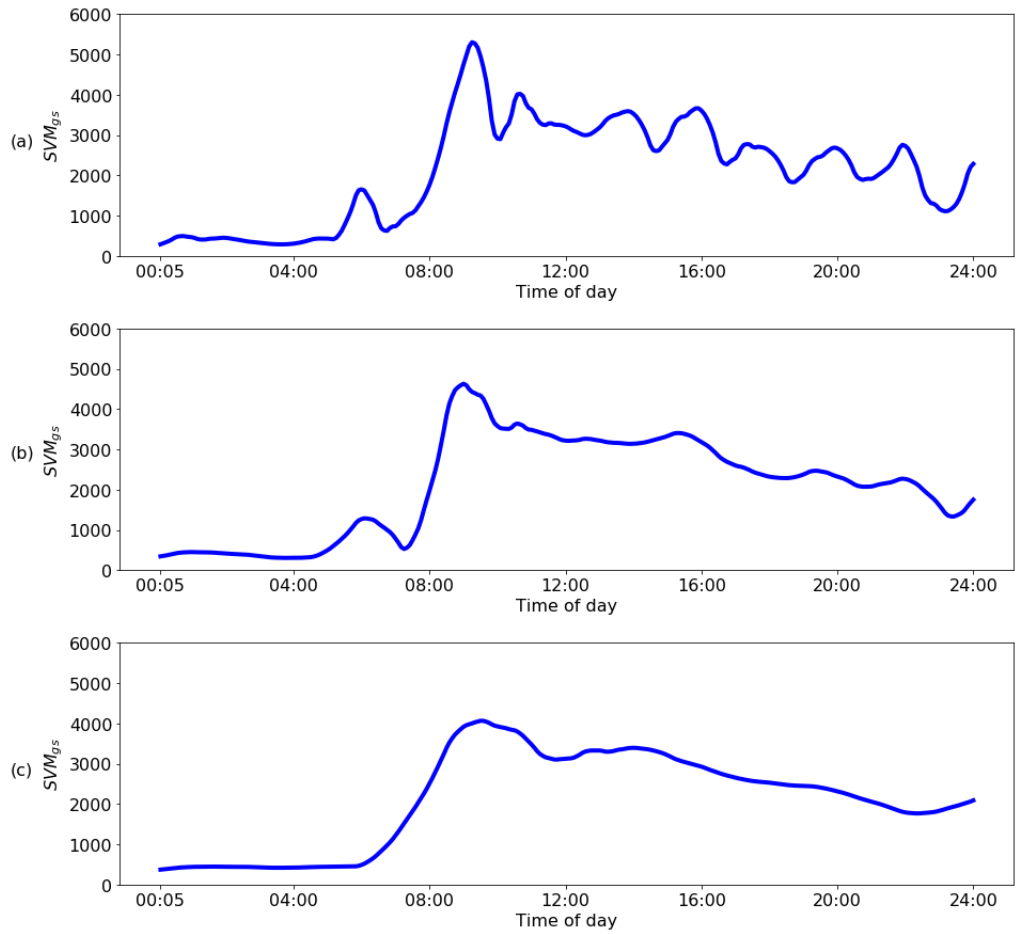


Figure 5.8: ID:1143 scale approximations for 5 minute epoch. (a) 3th, (b) 4th, (c) 5th

The scale of the y-axis in Figure 5.8 is much greater than previous figures displaying activity profiles as more data is now being collapsed into an epoch. Once more, FPCA can be run for each of these before clustering. The concordance to our original solution are shown in Table 5.10.

Scale approximation	3rd	4th	5th
%	99	98	85

Table 5.10: Concordance percentages: 5 minute epoch with 3rd, 4th and 5th scale approximations

The 5^{th} scale approximation has a much lower concordance than the other two. This can be explained by looking at Figure 5.8(c), where the profile lacks much of the detail than the others possess. The 3^{rd} scale approximation has a slightly better level of agreement than the 4^{th}, so it will be used to explore the differences further. The IDs for the individuals that changed cluster are displayed in Table 5.11.

ID	1 minute	5 minute
1743	High	Morning
2676	Moderate	High

Table 5.11: Cluster change IDs: 1 minute vs. 5 minute epochs

Once again ID 2676 moves from the moderate to the high cluster, as it has done in each section thus far. It has been highlighted as a borderline case, and perhaps it belongs in a different cluster. Nonetheless a concordance of 99% means that our clusters are not sensitive between an epoch length of either 1 minute or 5 minutes.

Another choice made in the data preprocessing was to use only the weekday, Monday to Friday, data and not include the weekend data despite each individual wearing the accelerometer for a week. In the section 5.4, the effects of including the weekend data will be explored.

5.4 Weekday vs weekend

Studies have documented that four to seven days monitoring may be needed to obtain reliable information on habitual physical activity (Matthews, Ainsworth, Thompson, & Bassett, 2002; Dillon et al., 2016). To illustrate the effect of including the weekend, the profiles for our example individual, ID 1143, are shown in Figure 5.9.

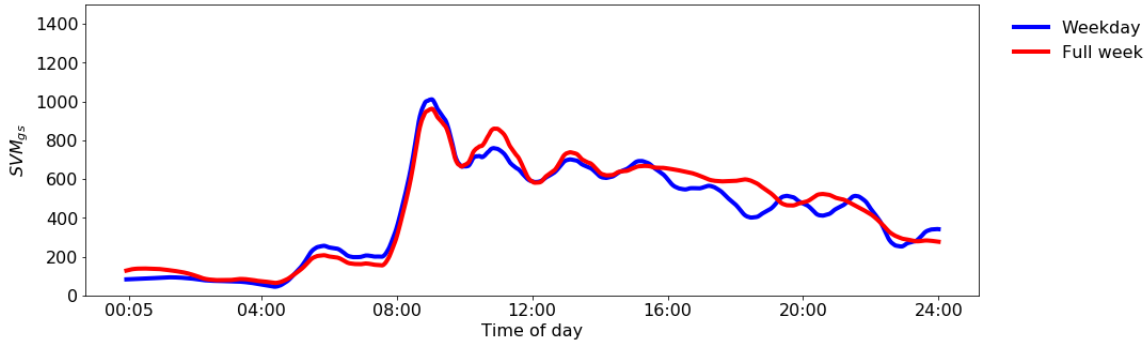


Figure 5.9: Profiles for ID 1143 using data from the weekdays and full week

This individual is more active in the afternoon at the weekends and so the average profile is brought up around this time. Overall, the profile is largely similar. However when clustering is performed the concordance between the two methods is only 64%, with 138 people being clustered differently. What this suggests is that the inclusion of the weekend data has had a large effect on the profiles of people in the cohort. To investigate these differences further, the cluster centroids for each solution are displayed in Figure 5.10.

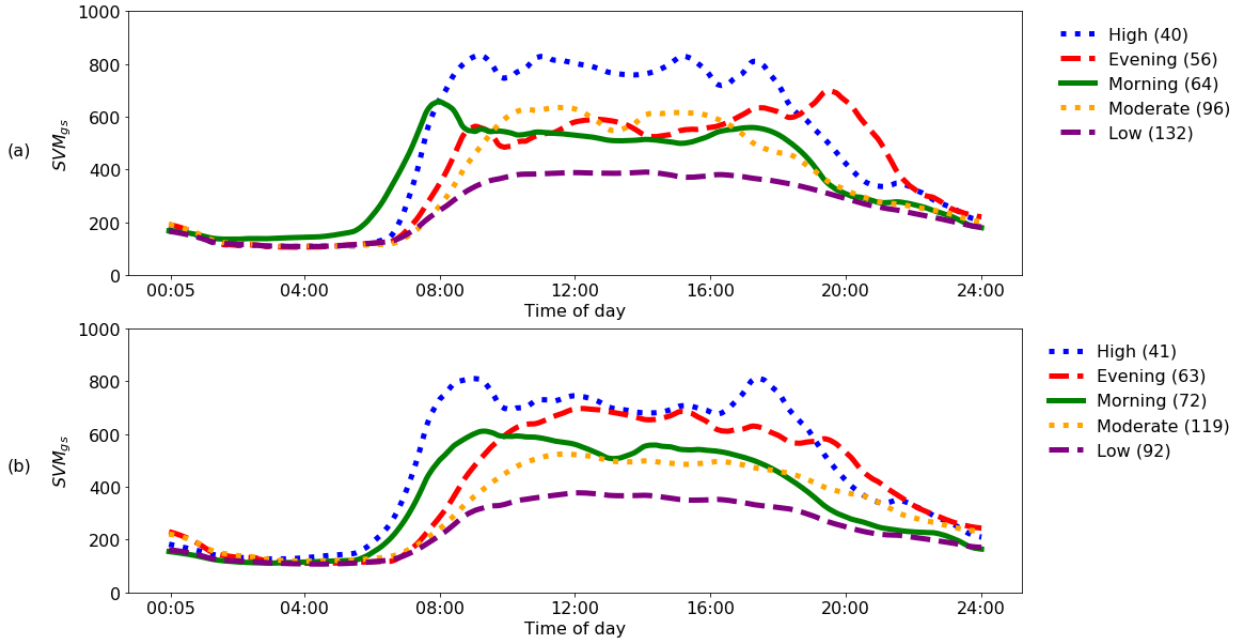


Figure 5.10: Cluster centroids. (a) Weekday only, (b) Full week

The characteristics of the full week centroids, which is the average profile for those in the cluster, is quite different to those that only used the week day data. When activity profiles vary with day of the week, information is lost when you average. As such, given that profiles differ on weekends it is not a good idea to create profiles by averaging over 7 days. Clustering could be performed based solely on the weekend data, but that will not be considered in this thesis.

Having concluded the robustness of the cluster solution with respect to certain characteristics, the final chapter will provide a discussion of the findings.

Chapter 6 - Discussion

In this chapter, a discussion of the findings is presented. The limitations of the analysis are then discussed along with recommendations. The section on implications examines translating research into practice before a conclusion is presented.

6.1 Discussion of findings

The first challenge in this study was to convert the raw accelerometer data into an activity profile in order to compare individuals. From an initial exploration of the data, it was evident that sharp discontinuities were present and so it was deemed that wavelets would be a good choice of smoothing technique. In the sensitivity analysis, it was found that wavelets performed similarly to cubic splines with uniformly placed knots at hourly intervals.

The goal of the cluster analysis was to explore if distinct groups existed in the cohort based solely on an objective measure of their physical activity. A combination of hierarchical and non-hierarchical (K-means) methods were used to determine the appropriate number of clusters. From the analysis performed, the five cluster solution was deemed the most appropriate. These clusters were labelled as high, moderate, low, evening and morning. Two distinct clustering methods produced the exact same clusters of people.

The first method used just the distance between profiles, more specifically it was the sum of the Euclidean distances at each time point along the profile that was used as the metric for putting people into clusters. The second method utilised functional principal component analysis to decompose all the curves into their dominant modes of variation being using the respective component scores to cluster. The most dominant pattern, the first principal component, had a very similar shape to the average curve. Which meant that differences in the averages dominated everything and was the main source of variation amongst people.

Profiling of the clusters revealed a divide by activity patterns, perhaps it is suggested by chronotype. Like other personality traits, chronotype extends along a continuum, with a few extremes at each end, and most people clustering in the middle. The distribution resembles a normal distribution from morningness to eveningness (Randler, 2009). Individuals in the tails of the distribution are colloquially known as morning "larks" and night "owls", reflecting that they either go to sleep early and wake early, or go to sleep late and wake late. If you are a morning lark or a night owl, then it is likely that you are fully aware of it. If you do not have a preference for morning, afternoon, or evening, then you are a neutral (neither) chronotype. Determining your

chronotype is currently done using a series of questions, but perhaps accelerometers can be used in conjunction with these questions.

Knowing your chronotype will allow you to work with your body rather than against it. A growing number of companies are encouraging their employees to work when their bodies are most awake by offering flexible hours. The Society for Human Resource Management (SHRM) conducted a survey in 2018 that found that 57% of its members offer flexible working hours, which was 5% more than in 2014 (SHRM, 2018). The result being a well rested and more productive work force. A real world experiment was conducted in a steel factory in Germany, where they aligned work schedules with chronotypes (Vetter, Fischer, Matera, & Roenneberg, 2015). Day shifts were assigned to the morning larks and the night shifts to the night owls. By aligning schedules with internal clocks, researchers found that employees got 16% more sleep. The benefits of getting sufficient levels of sleep are well documented.

Everyone has an optimal time to fall asleep and wake up. The quality of sleep decreases when you do not sleep when your body wants to, leading to fatigue and health problems. If you require a alarm clock to wake up in the morning then you are out of sync with your internal rhythm. Worker fatigue has been responsible for many work place accidents. One famous example was NASA's Space Shuttle Challenger which broke apart 73 seconds into flight killing its entire crew. The main contribution for the incident was human error and poor judgement related to sleep loss and shift work during the early morning hours (Feynman, 1986).

Unfortunately the chronotypes for the individuals is not available, so the suggestion of identifying morning and evening people can not be validated. In the next section further limitations will be discussed.

6.2 Limitations and Recommendations

This was a cross-sectional study, conducted in one primary care centre in 2011 with a sample of people aged 50-69, which limits the generalisation of our findings to other populations. However, a follow up study was conducted for this cohort in 2016. So baseline activity can be related to follow up. Also changes in these profiles could be investigated. This follow up data is, however, out of the scope of this study.

People can not be forced to wear the accelerometers. If the most inactive people simply refused to wear the accelerometers, it could skew the results. This introduces the potential for bias. In this study, valid accelerometer data was only available for 397 people from the 2047 in the cohort. The activity profiles of those included in the study may not be typical of the activity profiles of the

population of interest. Therefore, it is possible for example that those who refused to participate had an activity profile that was not captured by the profiles presented in this thesis. As such, my results may not present a complete picture.

To investigate whether the 397 is representative sub-sample of the cohort, the distributions between those with accelerometer data and the full cohort can be investigated in relation to certain demographic information. These are shown in Appendix A, for age, BMI, education and smoking status. They illustrate that the sub-sample is representative of the cohort.

Another consideration is that people may be more active than usual because they know they are being monitored. This is known in behavioural psychology as the Hawthorne effect (Parsons, 1974), which is a type of reactivity in which individuals modify an aspect of their behaviour in response to be being observed or being part of a study (Rowland, 1994). People can not maintain this posturing for long and so the effects diminish over time (Leonard & Masatu, 2006). A longer time frame for the study could be recommended in order to lessen this effect.

The feasibility of conducting a study similar to this on a larger scale is questionable given the logistics and cost of getting the devices to the participants. However, with technological advances, this is now possible, as shown by the UK Biobank study who collected and analysed accelerometer data for 100,000 participants (Doherty et al., 2017). In this study participants were contacted via email and the accelerometers were in turn posted to them with instructions on their use. Without physically having to report to a doctor's office, it was possible to collect such a large amount of activity data. The author then reported on activity variation by age, sex and other demographics. They had a response rate of 44.8% and 96,600 participants (93.3%) provided valid accelerometer data. If their approach to data collection was adopted, studies could potentially be conducted for a larger cohort and over a longer time frame.

In addition to pure accelerometers, new fitness trackers and smart watches are released every year, which contain accelerometers in addition to other sensors. Smart watches monitor heart rate in order to distinguish between just movement and actual exercise. Heart rate increases when exercising (Achten & Jeukendrup, 2003). A smart watch will only register movement as exercise once the heart rate goes above a certain threshold. Therefore using heart rate readings in conjunction with the movement data, leads to a more accurate estimate of METs or calories burned. Walking at a leisurely pace may not raise heart rate sufficiently to be considered exercise.

Using 1 minute epochs may obscure short bursts of VPA or MVPA and

underestimate high intensity PA (Welk, 2002; Nilsson, Ekelund, Yngve, & Söström, 2002). Traditionally 1 minute epochs were used due to limited memory capacity issues (Allison, 1995). For future work, shorter epochs could be explored as they would provide more detailed information about the intensity and duration of activity (Trost et al., 2001).

The metric used to collapse the raw data into these epochs could also be explored. In this study, the mean was used to aggregate the daily PA values, whereas the median or certain quantiles could potentially have been used. This leads to another limitation in this study, which is the choice of statistical techniques. Both smoothing and cluster analysis are open to interpretation (Rousseeuw, 1987). Each profile consisted of 1440 minute by minute measurements of PA which ranged from midnight to midnight. These discrete measurements were converted into a smooth curve. The sensitivity analysis in Chapter 5 demonstrated that the clustering was robust to the choice of smoothing technique. For K-means clustering, the first challenge is to find the correct number value of K (the number of clusters). Graphical aids, such as silhouette analysis, were used in this study to interpret the different cluster solutions. To ensure stability and robustness of the results, the algorithm was run 10000 times with different centroid seeds.

The tendency to be more active in the morning may just be because of work commitments or personal schedules. After a lifetime of balancing work and social commitments people do not know what their natural rhythm is. In an ideal study, people would have an extended period free with no morning or evening commitments. In addition, the people in the study would avoid caffeine, other stimulants and artificial light which can push your chronotype later. What time would these people then tend to fall asleep and wake up?

6.3 Implications

Other studies have clustered accelerometer data in terms absolute activity volume, as in high or low activity groups (Lee et al., 2013; Rovniak et al., 2010; Fairclough, Beighle, Erwin, & Ridgers, 2012). However they do not place too much value in determining what time of day people are active. The reason for this is that for the most inactive or sedentary people, getting any amount of exercise is the most important thing regardless of the time of day. Only 22.9% of U.S. adults met PA guidelines between 2010 and 2015 (Blackwell & Clarke, 2018).

The implication for those in our cohort that are in the low activity cluster, is that they are not gaining any of the health benefits from exercise. The goal for the individuals in this cluster would be to undertake any amount of exercise whenever they can. Tips for slightly increasing activity include getting off the

bus one stop earlier, taking the stairs instead of the elevator or escalator and parking as far away from the store as possible. For those in the moderate activity group, the aim would be to increase activity levels further. While the high activity individuals should strive to maintain their activity levels.

This study went a step further and identified the times of day people are active. Equipped with this information, more targeted interventions can be deployed for increasing PA. Many gyms offer early morning classes which would be suitable for those identified as being more active around this time. Evening walking groups can be directed towards those individuals in the respective cluster. Using the FPCA method the scores on the second and third component would only be needed to determine the targets for each intervention.

Further to targeted interventions, being able to determine whether people have a propensity to be active in either the morning or evening would be invaluable to both the individual and any potential employers. This has implications for both personal performance and well being. People know what parts of the day they have the most energy and when they are peaking mentally. Knowing this means that the day can be tailored accordingly, using lower energy periods for more mundane tasks while saving peak energy periods for more creative and demanding work.

By offering flexible working hours, companies are beginning to recognise that people have different chronotypes. Further to this, it is being proposed that teams could potentially be built around people's chronotypes. In a paper called "Chronotype diversity in teams", Volk et al. (2017) give examples of jobs that require sustained attention over time. These include the police, nurses and surveillance teams who could benefit from having a mix of early and evening types to ensure that there is always someone who is alert and engaged.

6.4 Conclusions

At the outset of this study, the aim was to identify and characterise individuals in a cohort based solely on their activity profiles. The first step was to convert the raw accelerometer data into a profile that reflected the underlying functional nature of a person's activity.

FPCA was applied to the data to uncover associations related to the time of day differences in activity. Using FPCA meant that a huge amount of data is reduced down to just a few subject specific scores. Two distinct clustering methods identified the exact same 5 subgroups in the cohort. These results were subject to a sensitivity analysis which ensured their robustness.

In addition to separating in terms of absolute activity, two groups showed a

tendency to being more active in either the earlier or later parts of the day. Admittedly this may be explained by people simply having different work or personal schedules but perhaps it is revealing an underlying physiological aspect of the individual that warrants further investigation.

Emerging research reveals that everyone has an optimal time to both fall asleep and wake up. This biological rhythm inherent in everyone is known as a chronotype. A person's chronotype describes their propensity for sleep and activity at particular times during a 24 hour period. If a person does not sleep when their body wants, the quality and duration of the sleep is affected. This leads to fatigue, poor work performance and also health problems (Chervin, 2000; Hafner, Stepanek, Taylor, Troxel, & van Stolk, 2017)

Chronotype is typically determined via questionnaire, examples include the morningness-eveningness questionnaire (MEQ) (J. A. Horne & Östberg, 1976) and the Munich ChronoType Questionnaire (MCTQ) (Roenneberg, Wirz-Justice, & Merrow, 2003). These questionnaires have the same drawbacks as any in that they are subjective and have the potential to be biased based on an individual's own perceptions of themselves. In conjunction with these questionnaires, I believe the objectivity of an accelerometer could be an invaluable feature in the calculation of an individual's chronotype.

This study attempted to identify sub groups in a cohort based solely on their activity data. A method which capitalised on the longitudinal nature of the data was deployed rather than a summary which would have masked this temporal affect. By adopting this approach sub groups with an inclination towards being active at certain times of the day were revealed.

References

Abbott, R., & Davies, P. (2004). Habitual physical activity and physical activity intensity: their relation to body composition in 5.0–10.5-y-old children. *European journal of clinical nutrition*, *58*(2), 285.

Achten, J., & Jeukendrup, A. E. (2003). Heart rate monitoring. *Sports medicine*, *33*(7), 517–538.

Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. *Algorithms and Application, Boca Raton: CRC Press*.

Ainsworth, B. E., Haskell, W. L., Herrmann, S. D., Meckes, N., Bassett Jr, D. R., Tudor-Locke, C., ... Leon, A. S. (2011). 2011 compendium of physical activities: a second update of codes and met values. *Medicine & science in sports & exercise*, *43*(8), 1575–1581.

Ainsworth, B. E., Haskell, W. L., Whitt, M. C., Irwin, M. L., Swartz, A. M., Strath, S. J., ... others (2000). Compendium of physical activities: an update of activity codes and met intensities. *Medicine and science in sports and exercise*, *32*(9; SUPP/1), S498–S504.

Allison, D. B. (1995). *Handbook of assessment methods for eating behaviors and weight-related problems: Measures, theory, and research.* Sage Publications, Inc.

Anton, H., & Rorres, C. (2010). *Elementary linear algebra: applications version.* John Wiley & Sons.

Bames, J., Behrens, T. K., Benden, M. E., Biddle, S., Bond, D., Brassard, P., ... others (2012). Letter to the editor: Standardized use of the terms" sedentary" and" sedentary behaviours". *Applied Physiology Nutrition and Metabolism-Physiologie Appliquee Nutrition Et Metabolisme*, *37*, 540–542.

Bellman, R. E. (2015). *Adaptive control processes: a guided tour* (Vol. 2045). Princeton university press.

Bholowalia, P., & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*(9).

Blackwell, D. L., & Clarke, T. C. (2018). State variation in meeting the 2008 federal guidelines for both aerobic and muscle-strengthening activities through leisure-time physical activity among adults aged 18-64: United states, 2010-2015. *National health statistics reports*(112), 1–22.

Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports*, *100*(2), 126.

Castillo-Retamal, M., & Hinckson, E. A. (2011). Measuring physical activity and sedentary behaviour at work: a review. *Work*, *40*(4), 345–357.

Chambers, J. M. (2017). *Graphical methods for data analysis: 0.* Chapman and Hall/CRC.

Chervin, R. D. (2000). Sleepiness, fatigue, tiredness, and lack of energy in obstructive sleep apnea. *Chest*, *118*(2), 372–379.

Coleman, K. J., Saelens, B. E., Wiedrich-Smith, M. D., Finn, J. D., & Epstein, L. H. (1997). Relationships between tritrac-r3d vectors, heart rate, and self-report in obese children. *Medicine and science in sports and exercise*, *29*(11), 1535–1542.

Cradock, A. L., Wiecha, J. L., Peterson, K. E., Sobol, A. M., Colditz, G. A., & Gortmaker, S. L. (2004). Youth recall and tritrac accelerometer estimates of physical activity levels. *Medicine and science in sports and exercise*, *36*(3), 525–532.

Dai, X., Hadjipantelis, P., Ji, H., Mueller, H., & Wang, J. (2017). *fdapace: Functional data analysis and empirical dynamics. r package version 0.4. 0.*

Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, *12*(1), 136–154.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., & De Boor, C. (1978). *A practical guide to splines* (Vol. 27). Springer-Verlag New York.

Dillon, C. B., Fitzgerald, A. P., Kearney, P. M., Perry, I. J., Rennie, K. L., Kozarski, R., & Phillips, C. M. (2016). Number of days required to estimate habitual activity using wrist-worn geneactiv accelerometer: a cross-sectional study. *PloS one*, *11*(5), e0109913.

Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., ... others (2017). Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study. *PloS one*, *12*(2), e0169649.

Draper, N. R., & Smith, H. (2014). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Eckel, R. H., Krauss, R. M., et al. (1998). American heart association call to action: obesity as a major risk factor for coronary heart disease. *Circulation*, *97*(21), 2099–2100.

Fairclough, S. J., Beighle, A., Erwin, H., & Ridgers, N. D. (2012). School day segmented physical activity patterns of high and low active children. *BMC public health*, *12*(1), 406.

Feynman, R. (1986). Report of the presidential commission on the space shuttle challenger accident. *Appendix F*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York, NY, USA:.

Gelman, A., & Imbens, G. (2018). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 1–10.

Graps, A. (1995). An introduction to wavelets. *IEEE computational science and engineering*, *2*(2), 50–61.

Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., & van Stolk, C. (2017). Why sleep matters—the economic costs of insufficient sleep: a cross-country comparative analysis. *Rand health quarterly*, *6*(4).

Hagströmer, M., Oja, P., & Sjöström, M. (2006). The international physical activity questionnaire (ipaq): a study of concurrent and construct validity. *Public health nutrition*, *9*(6), 755–762.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., et al. (2006). *Multivariate data analysis (vol. 6).* Upper Saddle River, NJ: Pearson Prentice Hall.

Hall, P., Müller, H.-G., Wang, J.-L., et al. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The annals of statistics*, *34*(3), 1493–1517.

Heath, G. W., Parra, D. C., Sarmiento, O. L., Andersen, L. B., Owen, N., Goenka, S., ... others (2012). Evidence-based intervention in physical activity: lessons from around the world. *The lancet*, *380*(9838), 272–281.

Horne, J., Brass, C., & Petitt, A. (1980). Circadian performance differences between morning and evening 'types'. *Ergonomics*, *23*(1), 29–36.

Horne, J. A., & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International journal of chronobiology*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90.

Ipsos, M., et al. (2016). *Healthy ireland survey 2015: summary of findings.* Department of Health (DoH).

Jacobs, J. D., Ainsworth, B. E., Hartman, T. J., & Leon, A. S. (1993). A simultaneous evaluation of 10 commonly used physical activity questionnaires. *Medicine and science in sports and exercise*, *25*(1), 81–91.

Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data.

Jones, E., Oliphant, T., & Peterson, P. (2014). {SciPy}: Open source scientific tools for {Python}.

Kangas, M., Konttila, A., Winblad, I., & Jamsa, T. (2007). Determination of simple thresholds for accelerometry-based parameters for fall detection. In *Engineering in medicine and biology society, 2007. embs 2007. 29th annual international conference of the ieee* (pp. 1367–1370).

Kearney, P. M., Harrington, J. M., Mc Carthy, V. J., Fitzgerald, A. P., & Perry, I. J. (2012). Cohort profile: the cork and kerry diabetes and heart disease study. *International journal of epidemiology*, *42*(5), 1253–1262.

Kerr, J., Marinac, C. R., Ellis, K., Godbole, S., Hipp, A., Glanz, K., ... Berrigan, D. (2017). Comparison of accelerometry methods for estimating physical activity. *Medicine and science in sports and exercise*, *49*(3), 617.

Lagerros, Y. T., & Lagiou, P. (2007). Assessment of physical activity and energy expenditure in epidemiological research of chronic diseases. *European journal of epidemiology*, *22*(6), 353–362.

Lee, P. H., Yu, Y.-Y., McDowell, I., Leung, G. M., & Lam, T. (2013). A cluster analysis of patterns of objectively measured physical activity in hong kong. *Public health nutrition*, *16*(8), 1436–1444.

Leonard, K., & Masatu, M. C. (2006). Outpatient process quality evaluation and the hawthorne effect. *Social science & medicine*, *63*(9), 2330–2340.

Lyche, T., & Morken, K. (2008). Spline methods draft. *University of Oslo*, *226*, 12.

Mannini, A., Intille, S. S., Rosenberger, M., Sabatini, A. M., & Haskell, W. (2013). Activity recognition using a single accelerometer placed at the wrist or ankle. *Medicine and science in sports and exercise*, *45*(11), 2193.

Matthew, C. E. (2005). Calibration of accelerometer output for adults. *Medicine and science in sports and exercise*, *37*(11 Suppl), S512–22.

Matthews, C. E., Ainsworth, B. E., Thompson, R. W., & Bassett, D. R. (2002). Sources of variance in daily physical activity levels as measured by an accelerometer. *Medicine and science in sports and exercise*, *34*(8), 1376–1381.

Matthews, C. E., & Freedson, P. S. (1995). Field trial of a three-dimensional activity monitor: comparison with self report. *Medicine and Science in Sports and Exercise*, *27*(7), 1071–1078.

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342.

Morgan, K., McGee, H., Dicker, P., Brugha, R., Ward, M., Shelley, E., ... others (2009). Slan 2007: Survey of lifestyle, attitudes and nutrition in ireland. alcohol use in ireland: A profile of drinking patterns and alcohol-related harm from slan 2007.

Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M., Herrick, R., & Gortmaker, S. L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association*, *101*(476), 1352–1364.

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 179–199.

Ngui, W. K., Leong, M. S., Hee, L. M., & Abdelrhman, A. M. (2013). Wavelet analysis: mother wavelet selection methods. In *Applied mechanics and materials* (Vol. 393, pp. 953–958).

Nilsson, A., Ekelund, U., Yngve, A., & Söström, M. (2002). Assessing physical activity among children with accelerometers using different time sampling intervals and placements. *Pediatric exercise science*, *14*(1), 87–96.

Nocon, M., Hiemann, T., Müller-Riemenschneider, F., Thalau, F., Roll, S., & Willich, S. N. (2008). Association of physical activity with all-cause and cardiovascular mortality: a systematic review and meta-analysis. *European Journal of Cardiovascular Prevention & Rehabilitation*, *15*(3), 239–246.

of Health, D., & Children, H. S. E. (2009). *The national guidelines on physical activity for ireland.* Department of Health and Children Dublin.

of Health, U. D., Services, H., et al. (2018). *2018 physical activity guidelines advisory committee scientific report.* Office of Disease Prevention and Health Promotion, Washington, DC.

Ohtake, Y., Belyaev, A., Alexa, M., Turk, G., & Seidel, H.-P. (2003). Multi-level partition of unity implicits. In *Acm transactions on graphics (tog)* (Vol. 22, pp. 463–470).

Organization, W. H. (2000). *Obesity: preventing and managing the global epidemic* (No. 894). Author.

Parsons, H. M. (1974). What happened at hawthorne?: New evidence suggests the hawthorne effect resulted from operant reinforcement contingencies. *Science*, *183*(4128), 922–932.

Pate, R. R., Pratt, M., Blair, S. N., Haskell, W. L., Macera, C. A., Bouchard, C., ... others (1995). Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine. *Jama*, *273*(5), 402–407.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Perry, I. J., Collins, A., Colwell, N., Creagh, D., Drew, C., Hinchion, R., & O'Halloran, T. D. (2002). Established cardiovascular disease and cvd risk factors in a primary care population of middle-aged irish men and women.

Radloff, L. S. (1977). The ces-d scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, *1*(3), 385–401.

Ramsay. (2005). Functional data analysis. *Encyclopedia of Statistics in Behavioral Science*.

Ramsay, & Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer.

Randler, C. (2009). Validation of the full and reduced composite scale of morningness. *Biological Rhythm Research*, *40*(5), 413–423.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.

Roenneberg, T., Wirz-Justice, A., & Merrow, M. (2003). Life between clocks: daily temporal patterns of human chronotypes. *Journal of biological rhythms*, *18*(1), 80–90.

Romesburg, C. (2004). *Cluster analysis for researchers*. Lulu. com.

Rossum, G. (1995). Python reference manual.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Rovniak, L. S., Sallis, J. F., Saelens, B. E., Frank, L. D., Marshall, S. J., Norman, G. J., ... Hovell, M. F. (2010). Adults' physical activity patterns across

life domains: Cluster analysis with replication. *Health Psychology*, *29*(5), 496.

Rowland, T. W. (1994). *On exercise physiology and the psyche.*

Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic journal of statistics*, *3*, 1193.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). Sensitivity analysis in practice: a guide to assessing scientific models. *Chichester, England*.

SHRM. (2018). 2018 employee benefits, the evolution of benefits. Retrieved from `https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Doc` `Employee Benefits Report.pdf`

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., & Freedson, P. (2009). An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology*, *107*(4), 1300–1307.

Strang, G., & Nguyen, T. (1996). *Wavelets and filter banks*. SIAM.

Sylvia, L. G., Bernstein, E. E., Hubbard, J. L., Keating, L., & Anderson, E. J. (2014). Practical guide to measuring physical activity. *Journal of the Academy of Nutrition and Dietetics*, *114*(2), 199–208.

Taillard, J., Philip, P., Chastang, J.-F., & Bioulac, B. (2004). Validation of horne and ostberg morningness-eveningness questionnaire in a middle-aged population of french workers. *Journal of biological rhythms*, *19*(1), 76–86.

Talbot, L. A., Gaines, J. M., Huynh, T. N., & Metter, E. J. (2003). A home-based pedometer-driven walking program to increase physical activity in older adults with osteoarthritis of the knee: a preliminary study. *Journal of the American Geriatrics Society*, *51*(3), 387–392.

Team, R. C., et al. (2013). R: A language and environment for statistical computing.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276.

Troiano, R. P., Berrigan, D., Dodd, K. W., Masse, L. C., Tilert, T., & McDowell, M. (2008). Physical activity in the united states measured by accelerometer. *Medicine & Science in Sports & Exercise*, *40*(1), 181–188.

Trost, S. G., Kerr, L., Ward, D. S., & Pate, R. R. (2001). Physical activity and determinants of physical activity in obese and non-obese children. *International journal of obesity*, *25*(6), 822.

Tucker, J. M., Welk, G. J., & Beyler, N. K. (2011). Physical activity in us adults: compliance with the physical activity guidelines for americans. *American journal of preventive medicine*, *40*(4), 454–461.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Ullah, S., & Finch, C. F. (2013, Mar 19). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, *13*(1), 43. Retrieved from `https://doi.org/10.1186/1471-2288-13-43` doi: 10.1186/1471-2288-13-43

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22.

Vetter, C., Fischer, D., Matera, J. L., & Roenneberg, T. (2015). Aligning work and circadian time in shift workers improves sleep and reduces circadian disruption. *Current Biology*, *25*(7), 907–911.

Walnut, D. F. (2013). *An introduction to wavelet analysis*. Springer Science & Business Media.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Washburn, R. A., & Montoye, H. J. (1986). The assessment of physical activity by questionnaire. *American Journal of Epidemiology*, *123*(4), 563–576.

Wasserman, L. (2007). *All of nonparametric statistics, 268 pp.* Springer, New York.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Waxman, A. (2004). Who global strategy on diet, physical activity and health. *Food and nutrition bulletin*, *25*(3), 292–302.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

Welk, G. (2002). *Physical activity assessments for health-related research*. Human Kinetics.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*(470), 577–590.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., Lanas, F., ... others (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the interheart study): case-control study. *The lancet*, *364*(9438), 937–952.

Appendices

Appendix A

This appendix contains histograms and bar charts to display. For the continuous variables, age and BMI, their respective histograms are shown in Figure A.1 and Figure A.2.

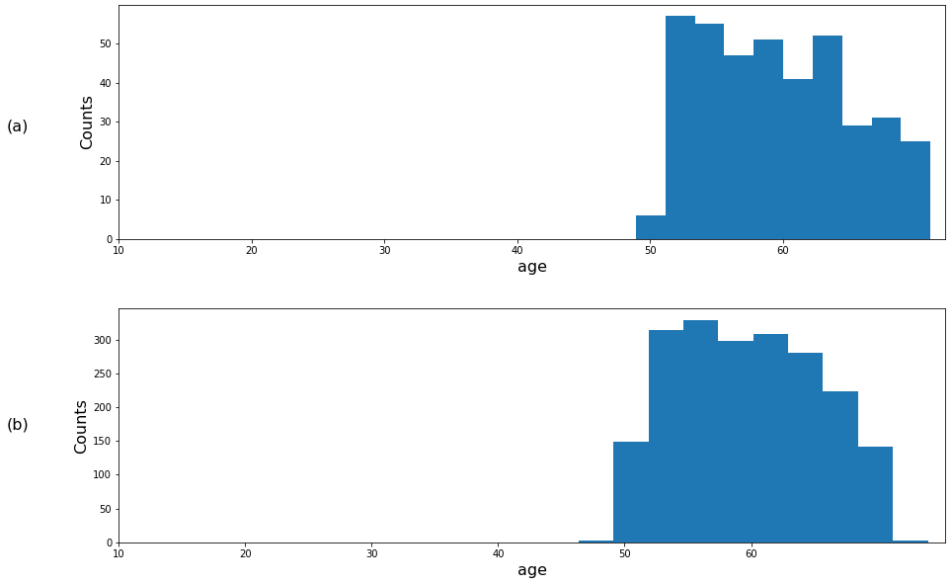


Figure A.1: Age distribution (a) Accelerometer group, (b) Full cohort

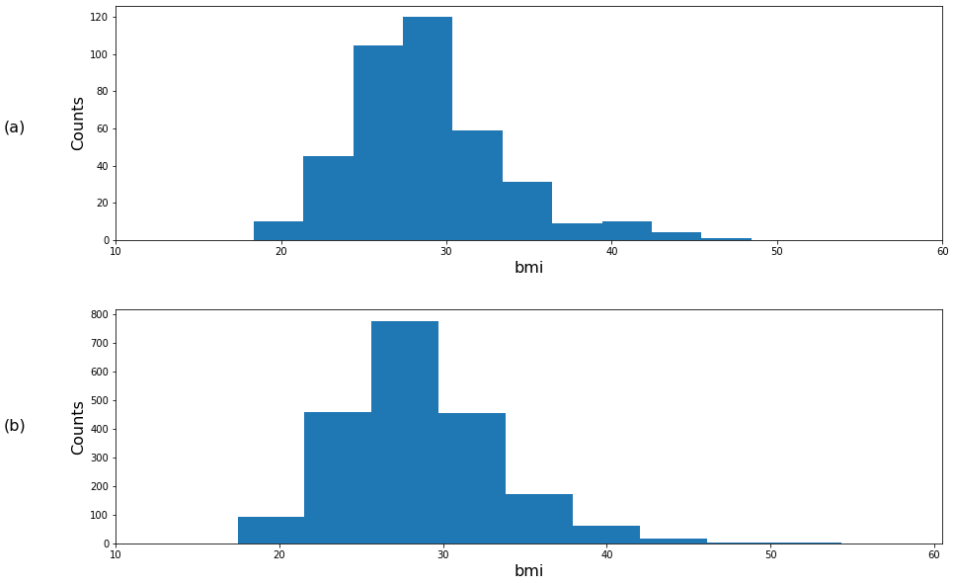


Figure A.2: BMI distribution (a) Accelerometer group, (b) Full cohort

For the categorical variables, education and smoking status, the bar charts representing the respective counts for each category are shown in Figure A.3 and Figure A.4.

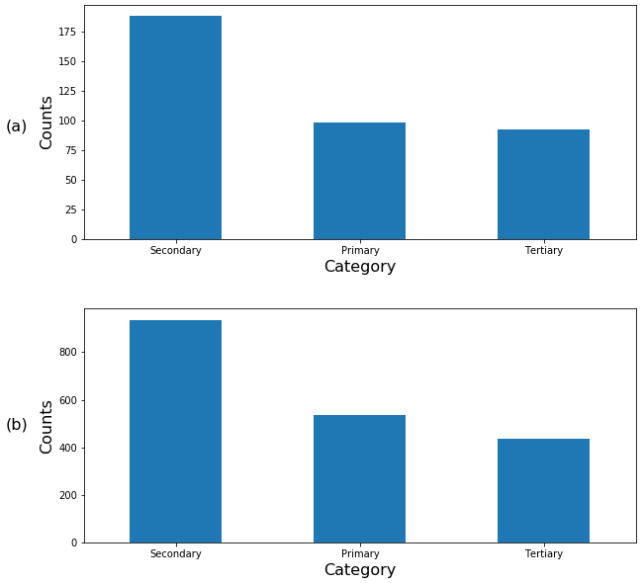


Figure A.3: Education distribution (a) Accelerometer group, (b) Full cohort

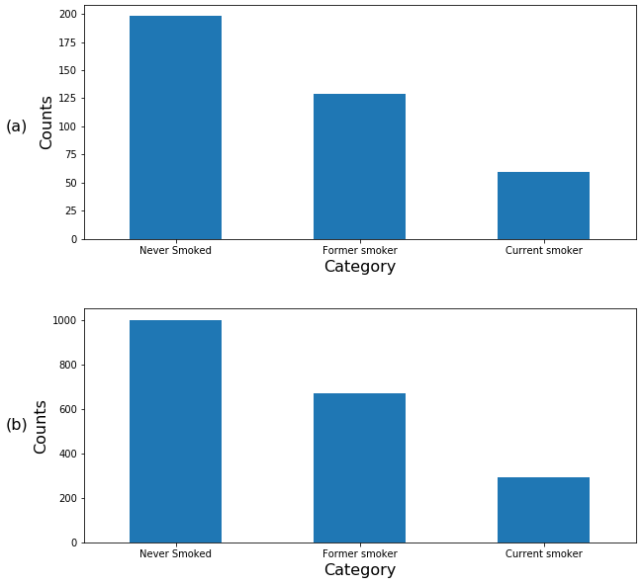


Figure A.4: Smoking status distribution (a) Accelerometer group, (b) Full cohort