

Inferring the Transportation Mode from Sparse GPS Data

Adel Bolbol

This thesis submitted for the degree of Doctor of Philosophy (PhD)

Department of Civil, Environmental and Geomatic Engineering
University College London

December 2013

Declaration

I, Adel Bolbol, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Understanding travel behaviour and travel demand is of constant importance to transportation communities and agencies in every country. Nowadays, attempts have been made to automatically infer the modes of transport from positional data (such as GPS data) to significantly reduce the cost in time and budget of conventional travel diary surveys. Some limitations, however, exist in the literature, in aspects of data collection (spatio-temporal sample distribution, duration of study, granularity of data, device type), data pre-processing (managing GPS errors, choice of modes, trip information generalisation, data labelling strategy), the classification method used and the choice of variables used for classification, track segmentation methods used (clustering techniques), and using transport network datasets. Therefore, this research attempts to fully understand these aspects and their effect on the process of inference of mode of transport. Furthermore, this research aims to solve a classification problem of sparse GPS data into different transportation modes (car, walk, cycle, underground, train and bus).

To address the data collection issues, we conduct studies that aim to identify a representative sample distribution, study duration, and data collection rate that best suits the purpose of this study. As for the data pre-processing issues, we standardise guidelines for managing GPS errors and the required level of detail of the collected trip information. We also develop an online WebGIS-based travel diary that allows users to view, edit, and validate their track information to assure obtaining high quality information. After addressing the validation issues, we develop an inference framework to detect the mode of transport from the collected data. We first study the variables that could contribute positively to this classification, and statistically quantify their discriminatory power using ANOVA analysis. We then introduce a novel approach to carry out this inference using a framework based on Support Vector Machines (SVMs) classification. The classification process is followed by a segmentation phase that identifies stops, change points and indoor activity in GPS tracks using an innovative trajectory clustering technique developed for this purpose. The final phase of the framework develops a network matching technique that verifies the classification and segmentation results by testing their obedience to rules and restrictions of different transport networks. The framework is tested using coarse-grained GPS data, which has been avoided in previous studies, achieving almost 90% accuracy with a Kappa statistic reflecting almost perfect agreement.

Acknowledgements

Many people inspired, helped, and supported me in the journey to produce the work done in this thesis.

This work would not have been possible without the guidance and help of my primary supervisor Prof Tao Cheng and her support since 2009. The endless discussions, critical comments, and encouragement by Tao have been the corner stone of progression in my research career, and therefore, I would like to sincerely thank her for her limitless effort and patience. I would also like to deeply thank Dr Claire Ellul, my secondary supervisor, who has also been my great mentor and guide through the academic ways and has helped me advance through my work with firm steps.

I would like also to immensely thank my industrial supervisor, Dr Chris Marshall from u-blox, with whom research discussions and even normal conversations were as insightful and stimulating as writing a prize-winning conference scientific paper.

The decision to pursue a PhD degree would have not appeared in the horizon of my aspirations without being inspired by great examples in my life such as my father Prof Saad Bolbol and my uncle Dr Adel Bolbol. My passion of GI research specifically has been triggered earlier on by people that I owe a lot respect and gratitude such as Prof Jo Wood and Prof Jason Dykes that mentored me during my MSc at City University in 2007.

I would also like to deeply thank my mother María Elena Fernández and my sister Kamila for always supporting me in decision to be physically distant to follow my career passions.

Above all, I would like to thank my close friends, which have always been my torch in the darkest of times. The long list starts with, and is not limited to, my colleagues/friends in the SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering at UCL, administrative personnel, and other close friends that are my family-for-life.

Definitions

Movement Terminologies

| | |
|------------------------|---|
| GPS fix: | A horizontal positioning capture by a GPS device holding Northing and Easting coordinates of a trajectory at a given point in time |
| GPS Segment: | The line segment that joins two consequent GPS fixes of a trajectory. A GPS segment record holds the distance and duration travelled between the two GPS fixes. |
| GPS Stage: | <p>A group of consequent segments of the same mode of transport. A stage is a component part of a trip using a single mode of transport between interchanges. Walking is counted as a separate mode, but walks within single premises or between platforms at interchange stations are not included.</p> <p>E.g. A trip of 3 stages comprising a walk stage from home to a bus stop, a bus stage to central London, and a further walk stage to a place of work.</p> |
| Trip (Journey): | A complete door-to-door movement by an individual to achieve a specific purpose (E.g. travelling from home to work). A trip consists of one or more stages. |
| Mode Share: | A single trip may use several methods or modes of transport, which divide the trip into its separate stages. In this way, trip rates can be analysed by trip main mode, based on distance: the main mode of a trip is the mode on which the greatest proportion of the total trip distance is travelled (TfL, 2009b). |
| Trip Purpose: | <p>The purpose of a trip is defined by the activity at the destination, except when the trip is returning home in which case the purpose is defined by the activity at the origin. The following purposes are defined:</p> <ul style="list-style-type: none"> • Work/commuting - travel to or from the respondent's usual place of work; • Employer's business/other work – travel in course of work or to work at a location that is not the respondent's usual workplace; • Education – travel to or from school, college or university; • Escort education – accompanying a child to or from school; • Shopping and personal business – including shopping and use of services such as hairdressers, dry-cleaners, doctors, dentists, banks, solicitors etc; • Leisure – travel to or from entertainment, sport or social activities; • Other (including escort) – all purposes not otherwise classified, including accompanying or meeting someone for purposes other than education. |

London's Administrative Areas (TfL, 2009b)

| | |
|------------------------|--|
| | <p>The map shows the Greater London area with various administrative boundaries. The Central London area is highlighted in red, the Inner London area in green, and the Outer London area in blue. The map includes labels for various locations such as Watford, Stoken New, Slough, Windsor, Staines, Epsom, Woking, Sevenoaks, and Gray. A scale bar indicates distances up to 20,000 meters, and a north arrow is present in the top left corner.</p> |
| Greater London: | The area consisting of the 32 London boroughs and the City of London, and administered by the Greater London Authority. |
| Central London: | The Greater London Conurbation Centre or Central Statistical Area – an area roughly rectangular in shape, bounded by Regent's Park to the north, Whitechapel to the east, Elephant & Castle and Vauxhall to the south, and Kensington Gardens to the west. It is a larger area than the Central London Congestion Charging Zone (excluding the Western Extension), and includes the inner ring road and Paddington, Marylebone, Euston and King's Cross rail stations. |
| Inner London: | City of London, and the London boroughs of Camden, Hackney, Hammersmith and Fulham, Haringey, Islington, Kensington and Chelsea, Lambeth, Lewisham, Newham, Southwark, Tower Hamlets, Wandsworth and Westminster. |
| Outer London: | The London boroughs of Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Croydon, Ealing, Enfield, Greenwich, Harrow, Havering, Hillingdon, Hounslow, Kingston upon Thames, Merton, Redbridge, Richmond upon Thames, Sutton and Waltham Forest. |

Weekday Time Periods

| | |
|-------------------------------|-----------------|
| AM peak (morning peak) | 07:00 to 10:00. |
| Inter-peak | 10:00 to 16:00. |
| PM peak (evening peak) | 16:00 to 19:00. |
| Evening | 19:00 to 22:00. |
| Night-time | 22:00 to 04:00. |
| Early am | 04:00 to 07:00. |

| Work Status | |
|---------------------------|---|
| Working full-time: | People in paid employment normally working for more than 30 hours a week. |
| Working part-time: | People in paid employment working for not more than 30 hours a week. |
| Self-employed: | Those who in their main employment work on their own account, whether or not they have any employees. |

Interchangeable Terms

Some terms might be used interchangeably in this research and in other research studies. These are as follows:

- **Velocity** (and Speed): The measurement of the rate and direction of change in the position of an object, and in the case of this research, a human trajectory.
- **Participant** (user, respondent and trajectory): The individuals participating in any kind of testing involved in this research or in any previous research mentioned in this research. The term trajectory however, shall be only used in the context of the participant's trace in space.
- **Mode of Transport** (travel mode, transportation mode, means of transport, transport modality and form of transport): is a general term for the different kinds of transport facilities that are often used to transport people or cargo.
- **Confusion matrix** (and confusion index): In the field of artificial intelligence, a confusion matrix is an accuracy measurement indicator, typically used in supervised learning (in unsupervised learning it is typically called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.
- **Epoch rate** (polling rate, rate of collection, frequency of collection and sampling rate): A specific instant in time. GPS carrier phase measurements are made at a given frequency (e.g. every 30 seconds).
- **Underground** (and Tube): Refers to the underground mode of transport (or as commonly known as Metro), and both terms are used interchangeably all through the thesis based on the context.

Contents

| | | |
|-------|--|----|
| 1 | Introduction..... | 22 |
| 1.1 | Background | 22 |
| 1.1.1 | Transport Policies, Travel Planning and Travel Surveys | 22 |
| 1.1.2 | Data Collection in Transport Studies..... | 23 |
| 1.1.3 | GPS-Based Travel Surveys..... | 24 |
| 1.1.4 | Limitations of Current Mode Detection Attempts | 26 |
| 1.2 | Research Aim & Objectives | 27 |
| 1.3 | General Thesis Research Structure | 28 |
| 2 | Literature Review: Understanding Travel Behaviour From GPS Data | 32 |
| 2.1 | Introduction | 32 |
| 2.1.1 | Effect of Validation Parameters on Accuracy | 34 |
| 2.1.2 | Effect of the Inference Techniques on Accuracy | 35 |
| 2.1.3 | Chapter's Content | 37 |
| 2.2 | GPS Data Collection..... | 39 |
| 2.2.1 | Travel Survey Definitions..... | 39 |
| 2.2.2 | Used Device and Study Area..... | 39 |
| 2.2.3 | Sample Spatial Size and Duration for Mode of Transport Inference Validation..... | 40 |
| 2.2.4 | Data Collection Rate (Temporal Granularity)..... | 41 |
| 2.2.5 | Summary of Data Collection Issues..... | 42 |
| 2.3 | GPS Data Pre-Processing..... | 43 |
| 2.3.1 | Dealing with GPS Limitations | 43 |
| 2.3.2 | Choice of Modes and Degree of Trip Information Generalisation..... | 46 |
| 2.3.3 | Data Logging Strategy..... | 47 |
| 2.3.4 | Summary of Data Pre-Processing Issues | 49 |
| 2.4 | Mode of Transport Classification Techniques..... | 51 |
| 2.4.1 | Variable Selection..... | 51 |
| 2.4.2 | Classification Limitations | 52 |
| 2.4.3 | Summary of Classification Technique limitations..... | 53 |
| 2.5 | Identifying Stops and Mode Switches..... | 54 |
| 2.5.1 | Identifying Clusters..... | 54 |
| 2.5.2 | Identifying Change Points | 56 |
| 2.5.3 | Identifying Static Indoor Activity | 57 |
| 2.5.4 | Discussion on Segmentation Issues..... | 59 |

| | | |
|-------|--|-----|
| 2.6 | Using GIS with Transport Network Datasets | 61 |
| 2.6.1 | Map Matching | 61 |
| 2.6.2 | Network Matching | 62 |
| 2.6.3 | Summary of Methods using GIS with Transport Network Datasets..... | 63 |
| 2.7 | Summary and Conclusions | 65 |
| 2.7.1 | Validation (Data-Related) Processes | 66 |
| 2.7.2 | Classification Method-Related Processes..... | 67 |
| 3 | Framework of the Methodology | 72 |
| 3.1 | Framework Description | 72 |
| 3.1.1 | Validation (Data-Related) Processes | 73 |
| 3.1.2 | Classification Method-Related Phases | 75 |
| 3.2 | Addressing Research Aim and Objectives | 81 |
| 3.3 | Detailed Thesis Research Methodological Structure | 82 |
| 4 | Data Collection | 88 |
| 4.1 | Positioning Device Type | 89 |
| 4.1.1 | Positioning Systems | 89 |
| 4.1.2 | Mobile Technology Functionality Modes | 90 |
| 4.1.3 | Mobile Devices for Mode of Transport Inference Studies..... | 91 |
| 4.2 | Study Area | 93 |
| 4.3 | Sample Collection Rate..... | 93 |
| 4.3.1 | Data..... | 94 |
| 4.3.2 | Testing Methodology | 95 |
| 4.3.3 | Results and Analysis..... | 99 |
| 4.3.4 | Experiment Conclusion..... | 104 |
| 4.4 | Sample Spatial, Modal and Temporal Distributions | 105 |
| 4.4.1 | Collected Datasets..... | 105 |
| 4.4.2 | GPS Data Sample Spatial, Modal and Temporal Distributions in the Study Area 105 | |
| 4.4.3 | Limitations of Sample Properties | 115 |
| 4.5 | Summary of data Collection Issues Addressed in this Chapter | 116 |
| 5 | Data Pre-Processing | 120 |
| 5.1 | Managing GPS Errors and Choice of Modes | 120 |
| 5.1.1 | Low Positional Accuracy | 121 |
| 5.1.2 | Signal Loss | 121 |

| | | |
|-------|--|-----|
| 5.1.3 | Selected Modes of Transport | 122 |
| 5.2 | Trip information Generalisation | 122 |
| 5.2.1 | Elements of a Track..... | 122 |
| 5.2.2 | Segmentation and Labelling | 124 |
| 5.3 | Data Labelling Strategy..... | 129 |
| 5.3.1 | Introduction into Non-Expert Online Travel Diaries..... | 129 |
| 5.3.2 | Travel Diaries | 130 |
| 5.3.3 | Recalling Spatial Information | 131 |
| 5.3.4 | Geo-Web 2.0..... | 135 |
| 5.3.5 | Intelligent Travel Diaries | 136 |
| 5.3.6 | Application Conclusions & Further Work | 140 |
| 5.3.7 | Application Testing & Implementation | 141 |
| 5.4 | Summary of Data Pre-Processing Strategy | 143 |
| 6 | Phase I: Mode Classification (Moving Window-Based Support Vector Classification (SVC)) | 146 |
| 6.1 | Introduction | 146 |
| 6.2 | Independent Variable Selection | 147 |
| 6.2.1 | ANOVA Test for Variable Selection | 148 |
| 6.3 | Transportation Mode Classification..... | 152 |
| 6.3.1 | Classification Limitations | 152 |
| 6.3.2 | Support Vector Machines (SVM) Classification and Model Selection..... | 153 |
| 6.4 | Window-Based SVM Classification | 155 |
| 6.4.1 | Multi-Segment Instance Classification | 155 |
| 6.4.2 | Moving Window SVM Classification..... | 157 |
| 6.5 | Verification by Initial Modal Segmentation | 158 |
| 6.6 | Classification Results | 159 |
| 6.6.1 | Type I and II Errors | 160 |
| 6.6.2 | Integration Results | 161 |
| 6.7 | Summary | 162 |
| 7 | Phase II: Segmentation (Identifying Stops & Change Points) | 166 |
| 7.1 | Detecting Stops | 167 |
| 7.1.1 | Statistic Evaluation | 167 |
| 7.1.2 | Spatio-Temporal Clustering Algorithm | 169 |
| 7.1.3 | Validation & Results..... | 171 |

| | | |
|-------|--|-----|
| 7.1.4 | Discussion of Results | 173 |
| 7.2 | Identifying Static Indoor Activity (Gaps)..... | 175 |
| 7.2.1 | Exact Location of Indoor Activity..... | 175 |
| 7.2.2 | Occurrence of Indoor Activity..... | 178 |
| 7.2.3 | Indoor Activity Summary | 179 |
| 7.3 | Identifying Change Points (Integrating Classification and Segmentation Results)... | 179 |
| 7.4 | Segmentation Conclusions | 181 |
| 8 | Phase III: Verification (Network Matching) | 184 |
| 8.1 | Why Matching to a Network? | 184 |
| 8.2 | Available Network Data..... | 185 |
| 8.3 | Underground (Tunnel) Tube Network Matching (Time-Distance Underground Travel Detection Algorithm)..... | 187 |
| 8.3.1 | Network..... | 187 |
| 8.3.2 | Principle..... | 189 |
| 8.3.3 | Constraints | 191 |
| 8.3.4 | Algorithm and Limitations | 192 |
| 8.3.5 | Matching Accuracy | 193 |
| 8.4 | Underground (Non-Tunnel) Tube Network Matching..... | 196 |
| 8.4.1 | Network..... | 196 |
| 8.4.2 | Principle..... | 197 |
| 8.4.3 | Constraints | 197 |
| 8.4.4 | Algorithm and Limitations | 198 |
| 8.4.5 | Matching Accuracy..... | 201 |
| 8.5 | Train Network Matching..... | 205 |
| 8.5.1 | Network..... | 205 |
| 8.5.2 | Constraints | 206 |
| 8.5.3 | Matching Accuracy | 206 |
| 8.6 | Bus Network Matching..... | 209 |
| 8.6.1 | Network..... | 209 |
| 8.6.2 | Constraints | 210 |
| 8.6.3 | Matching Accuracy..... | 210 |
| 8.7 | Modal Classification Verification for Public Transit Network Matching | 213 |
| 8.7.1 | Segment Modal Decision Making Phase: Segment-Wise Horizontal Reasoning... | 213 |
| 8.7.2 | Re-Segmentation Phase | 215 |

| | | |
|--------|---|-----|
| 8.7.3 | Sequence Adjustment Phase: Stage-Wise Vertical Reasoning | 216 |
| 8.8 | Network Matching Accuracy | 218 |
| 8.9 | Chapter Summary | 220 |
| 9 | Further Validation (Results & Limitations) | 222 |
| 9.1 | Validation Dataset | 222 |
| 9.1.1 | Data Specifications | 222 |
| 9.1.2 | Data Profile | 223 |
| 9.1.3 | Device Carrying Information | 224 |
| 9.1.4 | Access to transport means | 225 |
| 9.1.5 | Movement Statistics | 226 |
| 9.1.6 | Data Extent and Scope | 227 |
| 9.1.7 | Data Summary | 228 |
| 9.2 | Validation | 229 |
| 9.2.1 | Moving Window SVM-Based Classification Results | 229 |
| 9.2.2 | Segmentation/Spatio-Temporal Clustering (Stops Detection) Results | 231 |
| 9.2.3 | Network Matching Algorithm Results | 231 |
| 9.3 | Discussion | 232 |
| 9.3.1 | Enhancing Accuracy by using Network Matching | 232 |
| 9.3.2 | External Limitations | 233 |
| 9.4 | Summary | 234 |
| 10 | Conclusions & Further Work | 238 |
| 10.1 | Introduction | 238 |
| 10.2 | Findings & Theoretical Implication | 239 |
| 10.2.1 | Research Question 1 | 239 |
| 10.2.2 | Research Question 2 | 241 |
| 10.2.3 | Research Question 3 | 241 |
| 10.2.4 | Research Question 4 | 242 |
| 10.3 | Policy implication | 243 |
| 10.4 | Limitations and Recommendation for Further Research | 244 |
| 10.5 | Conclusion | 246 |
| | Bibliography | 249 |
| | Appendices | 258 |
| | Appendix A: Placement Interview | 258 |
| | Appendix B: Exit Interview | 264 |

| | |
|--|-----|
| Appendix C: Paper Travel Diary Log Template..... | 266 |
|--|-----|

List of Figures

| | |
|--|-----|
| Figure 1.1 Links between Transportation Modelling Types with Survey Types | 23 |
| Figure 1.2 An Example of a “Trip Data” Survey Sheet (Appendix C) | 24 |
| Figure 1.3 Aim of the Research (Input: Raw Sparse GPS Data - Output: Trip Details) | 27 |
| Figure 2.1 Limitation Types affecting the Accuracy achieved by Methods attempting to Infer Transport Mode from GPS Data | 33 |
| Figure 2.2 Phases affecting Calculated Accuracy of Methods attempting to infer Transport Mode from GPS Data | 34 |
| Figure 2.3 Factors affecting Validation of Transport Mode Detection Methods from GPS Data | 35 |
| Figure 2.4 Processes affecting Accuracy of Transport Mode Detection Methods from GPS Data | 36 |
| Figure 2.5 Trip Segmentation Elements based on Mode of Transport | 39 |
| Figure 2.6 Segmentation Processes leading to identifying different Breaks..... | 54 |
| Figure 2.7 Clustering of observed stops in trips (crosses) to unique activity locations (boxes) (Schönfelder & Samaga, 2003) | 55 |
| Figure 2.8 Illustration of a location clustering algorithm (Ashbrook & Starner, 2003) | 56 |
| Figure 2.9 Number of places found for varying Dwell Time values for data from (Ashbrook & Starner, 2003) | 59 |
| Figure 2.10 Issues affecting Accuracy of Methods detecting Transport Mode from GPS Data. 65 | |
| Figure 3.1 Phases of Thesis based on Limitations & Opportunities mentioned in Chapter 2 | 73 |
| Figure 3.2 Framework produced in this Research to infer Transport Mode from Sparse GPS Data..... | 76 |
| Figure 4.1 Data Collection Elements affecting Validation of Positional Survey Methods..... | 88 |
| Figure 4.2 Desirable Properties of Positioning Devices..... | 91 |
| Figure 4.3 GPS Tracker Devices used to collect Data for this Research..... | 92 |
| Figure 4.4 London & its Diverse Transport Networks | 93 |
| Figure 4.5 1 Second Dataset & actual Route taken | 94 |
| Figure 4.6 Missing Data due to Lack of Fixes in a Dataset..... | 95 |
| Figure 4.7 Attributes that form the Epoch Rate Comparison Test | 95 |
| Figure 4.8 Shortest Distance calculation Method..... | 96 |
| Figure 4.9 Route with Data from different Datasets near End Destination | 97 |
| Figure 4.10 Map Matching selected Road Links for 30 seconds Track– Dataset 2..... | 98 |
| Figure 4.11 Box-Plots illustrating Positional Error of the 11 thinned Datasets | 100 |
| Figure 4.12 Total Route Length calculated from Datasets Compared to Actual Length..... | 101 |
| Figure 4.13 Average Speed calculated from different Datasets compared to Truth Average Speed..... | 103 |
| Figure 4.14 Distances from Set of Last Points in Each Dataset to End Destination | 104 |
| Figure 4.15 General Distribution of GPS Validation Dataset in London..... | 106 |
| Figure 4.16 Validation Dataset Heatmaps of different Zones of London..... | 106 |
| Figure 4.17 Validation Dataset Distribution for TfL Daily Time Periods in London | 108 |
| Figure 4.18 Validation Dataset Distribution for TfL Daily Time Periods in Central London.... | 109 |
| Figure 4.19 Validation Dataset Heatmaps for different Times of Day in Central London | 109 |
| Figure 4.20 Validation Dataset Heatmaps for different Times of Day in Inner London..... | 110 |

| | |
|--|-----|
| Figure 4.21 Validation Dataset Heatmaps for different Times of Day in Outer London..... | 110 |
| Figure 4.22 Validation Dataset Distribution for different Modes in London | 112 |
| Figure 4.23 Validation Dataset Distribution for different Modes in Central London | 112 |
| Figure 4.24 Validation Dataset Heatmaps of different Modes in Central London..... | 113 |
| Figure 4.25 Validation Dataset Heatmaps of different Modes in Inner London | 113 |
| Figure 4.26 Validation Dataset Heatmaps of different Modes in Outer London | 114 |
| Figure 5.1 GPS Hierarchical Data Structure in this Research | 123 |
| Figure 5.2 GPS Trip Elements | 123 |
| Figure 5.3 A Trip Description..... | 124 |
| Figure 5.4 “Bus Stop” Scenario Example..... | 127 |
| Figure 5.5 Example of an “Appointment Waiting” Scenario | 127 |
| Figure 5.6 Example of a “Sitting Outdoors” Scenario | 128 |
| Figure 5.7 Information Flow in/out of the Human Brain | 132 |
| Figure 5.8 Trade-Off between Rate of Memory Loss & Privacy Invasion Sentiment | 137 |
| Figure 5.9 The System’s Data Flow..... | 138 |
| Figure 5.10 The Application’s Interface Showing Point Edit Functions | 140 |
| Figure 6.1 Box Plots for the Values of different Independent Variables | 149 |
| Figure 6.2 Division of Data into Equal-Sized Instances (Three in this Case) | 155 |
| Figure 6.3 SVM Classification Accuracies using Different Lengths of Data Instances..... | 156 |
| Figure 6.4 Moving Window Classifying each 3-Segment Instance moving segment-by-segment along the track..... | 158 |
| Figure 7.1 Box Plots for Speed (above) & Distance (below) Values | 168 |
| Figure 7.2 General Description of Clustering Algorithm..... | 170 |
| Figure 7.3 Description of the Developed Algorithm used in Clustering for Stops | 171 |
| Figure 7.4 Effect of using different Percentiles of Speed & Distance Thresholds on achieved Accuracy & Error | 173 |
| Figure 7.5 Merging Typical Stop Cluster occurrences within a walk into one Segment..... | 174 |
| Figure 7.6 Example of Lack of Fixes between two Stages leading to creation of a Stop..... | 176 |
| Figure 7.7 Example of Results of k-means Spatial Clustering Technique used to identify the Location of Indoor Activity Occurrences..... | 177 |
| Figure 7.8 Indoor Segmentation Test Confusion Matrix Results | 179 |
| Figure 7.9 Coincidence of identified Stops/Gaps with SVM Classification Results | 180 |
| Figure 8.1 Conceptual Idea of Network Matching | 185 |
| Figure 8.2 Different London Transportation Networks..... | 186 |
| Figure 8.3 TfL’s London Underground Network Data (Data provided by: TfL)..... | 188 |
| Figure 8.4 TfL’s London Underground Network & Stations (Data provided by: TfL)..... | 189 |
| Figure 8.5 Distance(a) & Speed(b) travelled for Segments of less & more than 2 Minutes in Underground Mode | 190 |
| Figure 8.6 Time-Distance Scatter Plot for Segments of Underground Mode..... | 191 |
| Figure 8.7 Description of Underground Tunnel Detection Algorithm | 193 |
| Figure 8.8 Underground Travel Detection Results for Pilot Dataset | 195 |
| Figure 8.9 Underground Travel Detection Results for Pilot Dataset (Smaller Scale) | 195 |
| Figure 8.10 Tube Network Matching constrained by Temporal & Distance Thresholds | 197 |
| Figure 8.11 London Underground Tube Sections appearing above Ground (TfL, 2012d)..... | 198 |
| Figure 8.12 Network Matching Algorithm Flow Chart | 199 |

| | |
|---|-----|
| Figure 8.13 SVM Classified Segments used for Public Transport Network Matching | 200 |
| Figure 8.14 SVM Over-Classified Segments reversed by the Algorithm to Walk Mode..... | 201 |
| Figure 8.15 Example of Network-Matched GPS Fixes to London Underground Network (District Line) | 201 |
| Figure 8.16 Example of Network-Matched GPS Fixes to London Underground Network for Repetitive Trips | 202 |
| Figure 8.17 UK National Rail Network (Data provided by: OpenStreetMap) | 205 |
| Figure 8.18 Example of Matching to UK National Rail Network..... | 207 |
| Figure 8.19 Another Example of Matching to UK National Rail Network | 207 |
| Figure 8.20 Example of Mismatching to UK National Rail Network due to absence of Route Information | 208 |
| Figure 8.21 London TfL Bus Network (Data provided by TfL) | 210 |
| Figure 8.22 Example of Network-Matched GPS Fixes to London Bus Network..... | 211 |
| Figure 8.23 Another Example of Network-Matched GPS Fixes to London Bus Network | 211 |
| Figure 8.24 Modal Shares of Daily Journey Stages in London, 2007 (Source: (TfL, 2009b)) | 214 |
| Figure 8.25 Amending Mixed Non-Walk Stages | 216 |
| Figure 8.26 Merging Similar Non-Walk Stages Separated by Short durations of Walk Stages | 217 |
| Figure 8.27 Dismissing Extra Over-Classified Walk Segments from Ends of Non-Walk Stages | 217 |
| Figure 8.28 Amending Non-Walk Stages that contain a Segment Classified as Underground Mode using the Time-Distance Algorithm all into Underground | 218 |
| Figure 9.1 Gender Split across different Age Groups | 223 |
| Figure 9.2 Occupation Nature Type of the Validation Dataset by Gender..... | 224 |
| Figure 9.3 Device Carrying Time for different Age Groups | 224 |
| Figure 9.4 Number of GPS Segments/Fixes Collected by Age Group | 225 |
| Figure 9.5 Participant Ownership of Access to different Transport Means by Age Group..... | 226 |
| Figure 9.6 Average Distance vs. Time between GPS Fixes for each Participant..... | 226 |
| Figure 9.7 Histogram of Frequency of Average Speeds among Participants | 227 |
| Figure 9.8 Geographic Extent of most of the Collected Data..... | 228 |

List of Tables

| | |
|--|-----|
| Table 2.1 Summary of Previous Studies' Accuracies, Sample Sizes & Durations | 40 |
| Table 2.2 Transition Matrix of Modes of Transport (Zheng, et al., 2008)..... | 57 |
| Table 4.1 Properties of different Sensor Technologies used for Positioning..... | 90 |
| Table 4.2 Datasets with Different Holding Positions | 95 |
| Table 4.3 Route Length & Average Speed from the 11 Datasets & Actual Route..... | 102 |
| Table 4.4 Start-End Trip Distance Results (m) | 103 |
| Table 4.5 Spatio-Temporal Data Distribution in London (Colour Intensity-Coded according to Distribution Share) | 107 |
| Table 4.6 Spatio-Temporal Data Distribution in London Excluding Night Time Period (Colour Intensity-Coded according to Distribution Share) | 107 |
| Table 4.7 Spatio-Modal Data Distribution in London (Colour Intensity-Coded according to Distribution Share) | 111 |
| Table 4.8 Spatio-Temporal Data Distribution in London excluding Stops & Walk Modes (Colour Intensity-Coded according to Distribution Share) | 111 |
| Table 4.9 Modal-Temporal Data Distribution in London (Colour Intensity-Coded according to Distribution Share) | 114 |
| Table 4.10 Modal-Temporal Data Distribution in London Excluding Stops (Colour Intensity-Coded according to Distribution Share) | 115 |
| Table 5.1 Rules for Stop Locations Based on Trip Purpose | 125 |
| Table 5.2 Decisions for Specific Labelling Scenarios | 126 |
| Table 6.1 Tests of Equality of Group Means | 150 |
| Table 6.2 Tests of Equality of Group Means Results using different Independent Variables between Car, Train & all other Transportation Modes as a Third Category | 151 |
| Table 6.3 Confusion Matrix for Classification of Instances of 3 Segments | 157 |
| Table 6.4 Transition Matrix between Modes showing Probabilities of different Modal Mixes occurring (%)..... | 158 |
| Table 6.5 Confusion Matrix of Moving Window Algorithm based on Speed | 159 |
| Table 6.6 Confusion Matrix of Moving Window Algorithm based on Acceleration..... | 159 |
| Table 6.7 Symmetric Measurements for Cohen's Kappa Values for Speed & Acceleration | 160 |
| Table 6.8 Acceleration & Speed Confusion Matrix Result Difference | 160 |
| Table 6.9 Acceleration & Speed Results Type I & II Error Difference | 161 |
| Table 6.10 Integrated Acceleration-Speed Moving Window Algorithm Confusion Matrix | 161 |
| Table 7.1 Statistical Calculations for Speeds of Stationary Segments..... | 168 |
| Table 7.2 Statistical Calculations for Distance of Stationary Segments..... | 169 |
| Table 7.3 Accuracy & Error Measures for using Different Speed & Distance Values | 172 |
| Table 7.4 Clustering Algorithm Performance Results using different Speed & Distance Threshold Values..... | 172 |
| Table 8.1 Time-Distance Underground Tunnel Travel Detection Method Confusion Matrix Results..... | 193 |
| Table 8.2 Example of Output File of Time-Distance Underground Travel Detection Algorithm | 194 |
| Table 8.3 Differences between Network Data provided by TfL & OSM | 196 |
| Table 8.4 Underground Network Matching of GPS Fixes Confusion Matrix Results | 203 |

| | |
|---|-----|
| Table 8.5 Underground Network Matching of GPS Fixes Confusion Matrix Results (Dismissing the Time-Distance Outcome Trips) | 203 |
| Table 8.6 Underground Network Matching of GPS Stages Confusion Matrix Results | 203 |
| Table 8.7 Underground Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results | 204 |
| Table 8.8 Underground Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results (Dismissing the Time-Distance Outcome Trips)..... | 204 |
| Table 8.9 Rail Network Matching of GPS Stages Confusion Matrix Results | 208 |
| Table 8.10 Rail Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results | 209 |
| Table 8.11 Example of the Output from the Bus Network Matching Process | 212 |
| Table 8.12 Bus Network Matching of GPS Stages Confusion Matrix Results..... | 212 |
| Table 8.13 Bus Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results | 213 |
| Table 8.14 Example of Network Matching Results (matching to Bus & Underground Networks) | 214 |
| Table 8.15 Decision Matrix of Final Classification Assignment Based on Network Matching Results..... | 215 |
| Table 8.16 Confusion Matrix of SVM Results Before Network Matching | 218 |
| Table 8.17 Confusion Matrix of SVM Results After Network Matching..... | 218 |
| Table 8.18 Before & after Network Matching Accuracy Difference expressed in Stages | 219 |
| Table 8.19 Type I & Type II Errors Difference of before & after Network Matching | 220 |
| Table 9.1 Classification Results from applying the Moving Window SVM-based Classification Algorithm to Speed Values | 229 |
| Table 9.2 Classification Results from applying the Moving Window SVM-based Classification Algorithm to Acceleration Values | 230 |
| Table 9.3 Integrating Classification Results from applying the Moving Window SVM-based Classification Algorithm to Speed & Acceleration Values..... | 230 |
| Table 9.4 Integrating SVM-based Classification & Stop Detection Results..... | 231 |
| Table 9.5 Integrated Results of Classification, Stop Detection & Network Matching..... | 231 |
| Table 9.6 Result Differences of before & after applying the Network Matching Process | 232 |
| Table 9.7 Reduction in Type I/II Errors as a result of using Network Matching..... | 233 |
| Table 9.8 External Transport Modes classified using SVM & Stop Detection..... | 234 |
| Table 9.9 External Transport Modes classified using SVM & Stops Detection followed by Network Matching | 234 |

Chapter 1

Introduction

1 INTRODUCTION

This chapter describes the problem of automatically detecting the mode of transport from sparse GPS data. We first describe the background and motivation behind carrying out this work. Then, we define the aim and objectives of this thesis, which is followed by a description of the general structure of the thesis describing different phases of this research.

1.1 Background

1.1.1 Transport Policies, Travel Planning and Travel Surveys

The exponential expansion and population growth of cities nowadays give rise to an increasing need to understand travel behaviour to solve ever-growing transport problems. Understanding travel behaviour is important for many applications such as studying tourist activity (Edwards et al., 2009) or the impact of a strike on transportation systems (Tsapakidis, et al., 2012). Therefore, transport studies have always searched for ways to record and collect data regarding people's daily movements. As a result, travel surveys have emerged and are considered to be one of the most important ways of obtaining the critical information needed for transportation planning and decision making. Not only do these surveys gather current information about the demographic, socioeconomic, and trip-making characteristics of individuals and households, but they are also used to enhance our understanding of travel in relation to the choice, location, and scheduling of daily activities. This enables us to enhance our travel forecasting methods and improve our ability to predict changes in daily travel patterns in response to current social and economic trends and new investments in transportation systems and services. These travel surveys also play a role in evaluating changes in transportation supply and regulation as they occur (Griffiths, et al., 2000).

For example, the National Travel Survey (NTS) in Britain has been running continuously since 1988 by the Department for Transport's (DfT's) collecting information on personal travel in Britain across all transport modes (NTS, 2011). The survey data could be used for many purposes:

- Calibrating and maintaining the National Transport Model and the associated National Trip End Model dataset and WebTAG. These are standard tools used by transport planners in developing and appraising business cases;
- Evaluating the potential distributional impact of transport policies (through qualitative data on the travel patterns of different social and economic groups), and
- Estimating the total number of road accidents in Great Britain, including those accidents that are not reported to the police.

As part of these surveys, the different transport modes used at each journey are collected leading to understanding the modal choice/split which is considered one of the main four elements influencing the transport planning process (Ortúzar & Willumsen, 2011). Moreover, governments usually constitute transport policy based on a large number of instruments at their disposal (Rodrigue, et al., 2006). Among these instruments are safety and operating standards such as speed limits and road width which can be very much based on analysis from modal choice data. Collecting quantitative data about modal split along the spatio-temporal domain and correlating it with other sources of data, such as accidents rate, highlights factual

issues that help direct policy makers to address the most pressing and important transport issues (Van den Bossche, et al., 2005).

1.1.2 Data Collection in Transport Studies

As illustrated in the previous section, transportation surveys are a very important source for many applications in the transportation studies arena. One of which is transportation modelling such as trip generation, trip distribution, modal split/direct demand, discrete choice modelling and route assignment. Conventionally, travel surveys are carried out very differently according to the type of model that will be used in the study. This will also define the data needed, and hence, the data collection method to be used. As such, there are numerous types of travel surveys. The surveys could be typified or categorised according to properties such as the scope or the nature of the survey. Among these different survey types, but not limited to, are household surveys, intercept surveys, traffic and person counts, origin-destination surveys and travel time surveys (Ortúzar & Willumsen, 2011). As illustrated in Figure 1.1, every survey type serves feeds input into building a model type or more. However, the mode of transport is always an important if not an essential attribute needed to be collected from all surveys.

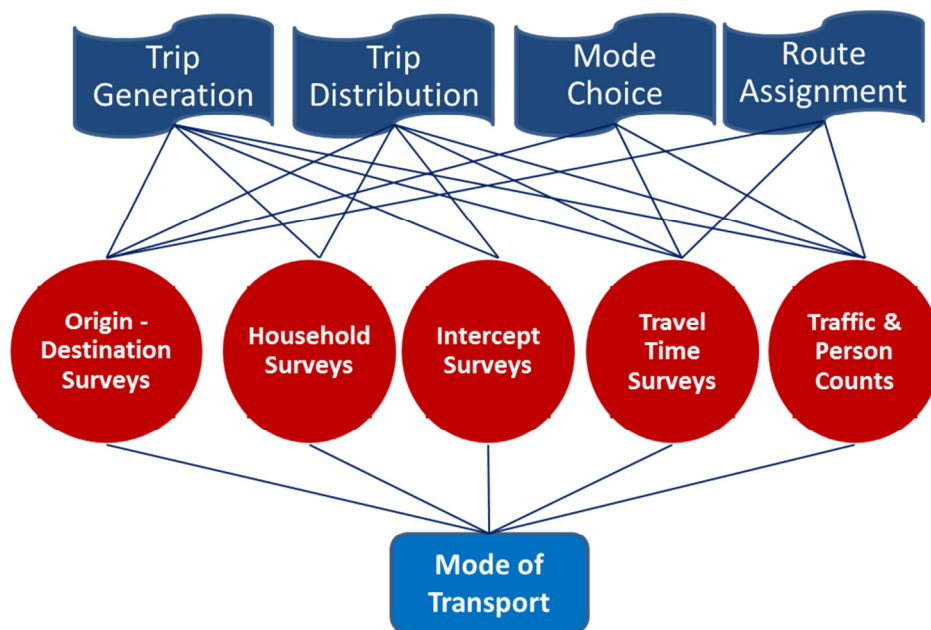


Figure 1.1 Links between Transportation Modelling Types with Survey Types

Most of the survey types are recommended to be collected split into two parts (1) personal and household survey and (2) trip data (Ortúzar & Willumsen, 2011). The former is concerned with the participant's details such as gender, age, possession of driving licence, etc. On the other hand, the latter part of the survey aims at detecting and characterising all the trips made by the participant. The attributes collected in this latter part include trips, stages (every change in travel mode), origin, destination, purpose, start and end times, mode used, etc. A typical example of a "trip data" survey sheet is illustrated in Figure 1.2.

| Name | | | Start Date: | | | | | |
|------|------|----|-------------------|---------|-----------|-------------------|-----------------|----------|
| Day | Time | | Duration (Min) | Purpose | Transport | Start Location | End Location | Comments |
| | From | To | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |

Figure 1.2 An Example of a “Trip Data” Survey Sheet (Appendix C)

This part of the survey has proven to be the most burdensome to the participants due to its day-to-day follow up nature, and hence results in low response rates. Moreover, many examples have proven to also be burdensome, dangerous, slow, difficult to plan, come back with inaccurate information and most of all expensive where some household surveys alone cost well over \$12 million to conduct which covered almost 200,000 respondents (Tierney, et al., 1996). The following provides a summary of the previous attempts to conduct different studies to collect travel data from either recruited or volunteered respondents. Along with that we shed some light on the various problems highlighted by different research.

Travel surveys held either by telephone or face-to-face is one of the oldest yet still common methods. This has proven to be a quite expensive, time consuming and sometimes dangerous to accomplish in some areas (Stopher & Metcalf, 1996). Other approach was to engage people by maintaining their travel diaries through the telephone, paper and pen, or computer-based (Stopher & Greaves, 2007). It was reported however that travel diary surveys have very low response rates because they were a burdensome task and some users omit certain travel information due to fear of their privacy invasion (Wolf, et al., 2003). A new trend emerged as to use GPS devices in conjunction with traditional surveys (Stopher, 2008). Some research even analysed GPS data in order to minimize trip under-reporting through improved survey methods (Bricka & Bhat, 2006).

1.1.3 GPS-Based Travel Surveys

More advanced methods started basing the travel diary entirely on GPS devices and subject the respondents to undertake prompted recall surveys. Some studies processed the user’s GPS data to provide maps to users of their tracks. Users were then asked to visually verify and validate any identified errors on the map (Stopher, 2008). Prompted recall surveys were considered the best means to obtain good accurate results, where respondents had the opportunity to see their tracks on a street map background which helped them identify errors and misinterpretations. However, among the main problems with prompted recall surveys are that they still achieved very low response rates (Stopher, 2008). Another problem was also the respondent’s accuracy in providing inputs because they were not very involved in the map making process was considered to be a specialist process.

More recently, new emerging technologies gave rise to using real time mobile technologies to collect location data and input respondent-entered data together in the same process. This seemed to be the most practical and time saving method that a travel diary could undertake. On the other hand, many problems emerged due to real time usage. Privacy issues, for example, arise unless there is a significant amount of time between when the data is collected and when the respondents fill in the surveys (Stopher, 2008). Another problem is the time lag between when users start on a new activity/mode and the when they record their status. Participants also seem reluctant to record everything they do due to the constant burden of having to have human interaction. Moreover, a common problem with current real time GPS devices is having a short battery life (GTrek, 2012). Finally, a very common problem in previous research is using incentives. A common practice is to pay participants to participate in surveys (Stopher & Metcalf, 1996), while some studies provided incentives to special access panels set up for conducting several different types of surveys on the same group of respondents (Roorda & Miller, 2004). A final common practice is to use volunteers that in many situations don't complete the full tasks (Stopher & Greaves, 2007).

It seems intuitive that by being able to automatically infer travel information from GPS tracks, such as modes of transport, can solve the problems identified above quite easily. Implementing this inference would potentially address most of the identified limitations in the previous section. These could be summarised into the following bullets.

- **Burden:** The user is relieved of having to fill a daily log during the whole duration of the study that in some studies could extend to more than several years.
- **Cost:** The expenses spent on recruiting people to link and verify user-generated information are minimised. The Impact on the survey cost as whole will be great due to reduced processing man-hours required and potentially decreased incentives offered to participants.
- **Slow:** The process of linking the participant-recorded trip data with the GPS data and verifying it will be reduced to a minimum (the time taken by the algorithm to infer this information).
- **Privacy:** The participants' privacy invasion is greatly reduced due to the fact that they do not need to report their whereabouts on a daily basis.
- **Accuracy:** Trip details no longer rely solely on the participant's memory.
- **User's Task Completion Time Gap:** Unlike real time self-reporting diaries, this process will not require participants to record trip information as soon as they end a stage of their trip.
- **User Incentive:** The participants will not require a bigger incentive to do a horrendous job, since they will be asked only to fill out a personal details form at the beginning of the survey and carry around a GPS device without having to worry about logging their day-to-day activities. Moreover, an incentive can even be providing participants with detailed information about their trips at the end of the survey period.

Therefore, much research is needed to automatically identify key trip information such as the mode of transport from GPS data. This not only requires a good understanding of the nature of multi-modal movement, but it also requires an extensive understanding of all the elements that affect the inference solution.

1.1.4 Limitations of Current Mode Detection Attempts

As a result of the research opportunity mentioned above, several research studies have emerged over the past decade aiming to infer the mode of transport from location information. Although these studies have proposed solutions to different problems that face this type of inference, yet many of these studies still suffered from several limitations. These limitations were related to either data quality or to the developed method of inference. **Data quality** limitations can fall into one of two categories; namely, data collection and pre-processing. These two groups of limitations could be further broken down into 5 issues. Data collection limitations include not accounting for sampling issues such as the sample spatio-temporal distribution, duration and temporal granularity. Moreover, many studies that collect positional data do not assess the suitability of the collection device or the study area. For data pre-processing limitations, however, include data validation issues such as the choice of modes to include, the level of information generalisation, the data labelling strategy, and decisions regarding the management of GPS errors.


On the other hand, **method-related** limitations include issues with classification methods used to separate modes, methods used for identifying stops in tracks leading to track segmentation, and attempts to use transport network datasets in the process. Regarding classification issues, most studies do not base the selection of variable to be used in the classification on any statistical analysis. Moreover, some of these studies use classification techniques that are not suitable for this type of inference. Whereas for the stop identification, many clustering methods do not account for slow speed modes such as walking nor accounts for situations of highly dense transport networks. Most of methods that cluster GPS data to identify stops also do not use labelled data to validate the developed algorithms. Moreover, very few of these methods perform the clustering in the spatiotemporal domain, instead focusing solely on the spatial domain to detect re-occurring stops. Another limitation is the immense lack of using transport network datasets to help inform mode detection algorithms, instead only using them to enhance the positional accuracy of GPS data.

Therefore, in this thesis we aim to address these limitations in order to standardise and enhance the accuracy of detecting the mode of transport from GPS data. The next section describes this aim in more detail, setting out the list of objectives that would help achieve this aim.

1.2 Research Aim & Objectives

As the previous section described, the research need of inferring trip information evolves from the necessity to obtain less problematic travel information. We have also highlighted the effect of attempting to develop a method that detects trip information such as the transport mode from GPS data on this research need. Therefore, the main aim of this work is to develop a method that automatically detects the hybrid mode of transport as a means of understanding human travel behaviour from sparse GPS data. The challenge is to produce a model that overcomes limitations and challenges of previous attempts which are briefly described in the next section and described in chapter 2 in detail. The developed method must also be robust enough to work with no information but raw GPS data in order to reduce participant burden. An example of the desired outcome is represented in Figure 1.3, and could be described as follows:

Given sparse GPS data as input, an output of the start (L_i) and end (L_{i+1}) locations of each trip from time t_i to t_k are identified along with the transport modes of each segment between t_i and t_k grouped into single mode stages (S_j to S_l) and any route details taken to do the trip.



| Time | Location | Stage | Mode | Segment | Route Info |
|-----------|-----------|-----------|------|-----------|------------|
| t_i | L_i | S_j | Walk | X_i | |
| t_{i+1} | | S_j | Walk | X_{i+1} | |
| t_{i+2} | | S_{j+1} | Bus | X_{i+2} | Bus (73N) |
| .. | | S_{j+1} | Bus | .. | Bus (73N) |
| .. | | S_{j+1} | Bus | .. | Bus (73N) |
| .. | | .. | Bus | .. | Bus (73N) |
| .. | | .. | Walk | .. | |
| t_k | L_{i+1} | S_l | Walk | X_k | |

Figure 1.3 Aim of the Research (Input: Raw Sparse GPS Data - Output: Trip Details)

This aim can only be understood and achieved by breaking down the main problem into several small counterparts. Addressing these smaller problems would formulate the objectives by which the main research problem is resolved. These research questions could be summarised as follows:

1. What best practices constitute the optimum standards of positional data collection and pre-processing which will ensure obtaining high quality data and a reliable performance validation process for movement-focused travel diaries?
2. What could constitute the characteristics of an ideal method to detect the transport mode of trip stages (as either stationary, walk, cycle, bus, car, train or underground)? And which independent variable(s) best discriminate between different classes (modes) in this classification problem?
3. Would breaking each trip into distinct stages, each consisting of a group of consequent GPS segments of the same mode of transport, enhance the accuracy achieved by this

classification? Could this also help identify stops and gaps within a GPS track; paving the way to further inferences?

4. What would be the effect of matching GPS fixes to their respective transport networks on such classification? And what type of information would such method require?

1.3 General Thesis Research Structure

As highlighted in this chapter, this thesis is dedicated to detecting the mode of transport from sparse GPS data. As described in subsection 1.1.4, this process consists of data and method-related limitations that can be further broken down into 5 areas of issues to be addressed along the process. Each of these areas counts as a step in the process of inference, and possesses several aspects that count as a research opportunity in their own respect. Some studies have attempted to address several of these aspects in different ways. **Chapter 2** describes these approaches along with a critical argument of benefits and limitations of each. In the process of presenting this argument, chapter 2 highlights the specific research gaps that need to be addressed in aspects of each stage of the inference process.

Consequently, **chapter 3** provides a full description of the framework with its 5 stages discussing how parts of each phase address certain limitations highlighted in light of chapter 2. Chapter 3 then describes how this framework addresses the aim and objectives specified earlier in this chapter. The final section of chapter 3 discusses the structure of the rest of the thesis based on the 5 stages of the developed framework along with the validation and discussion of results.

Chapters 4 and 5 describe the data-related stages of this research divided into data collection and data pre-processing respectively. These two chapters also describe our approach to address those issues that mainly affect the quality of the obtained positional data and related metadata describing the mode of transport. Data and metadata quality directly affects the performance of any proposed algorithm, as well as the validation instrument that assess its performance.

Chapters 6, 7 and 8 describe our approach to construct an inference framework to detect the mode of transport from sparse GPS data. These three chapters describe how our approach addresses the issues and limitations highlighted in chapter 2 regarding the classification, segmentation and using network datasets respectively. The core methodology of the framework lies in these three chapters, and therefore, we test the work developed in each chapter using a pilot dataset collected for the purpose of this research according to guidelines we previously set in chapters 4 and 5.

On the other hand, **chapter 9** re-tests the whole framework as a whole using the main dataset we collected for this research consisting of 95 participants for a period of 2 months for each participant. We approve of the suitability of the dataset by analysing participant information, participant movements, and the diversity of their modal mixes. We also discuss in chapter 9 the results obtained from applying the framework to the main dataset with no prior knowledge of participant information. Finally, **chapter 10** presents the conclusions of the work conducted in this research describing potential directions where future research can be carried out to build up on findings of this thesis.

Chapter 2

LITERATURE REVIEW: UNDERSTANDING TRAVEL BEHAVIOUR

2 LITERATURE REVIEW: UNDERSTANDING TRAVEL BEHAVIOUR FROM GPS DATA^{1 2}

2.1 Introduction

As discussed in chapter 1, understanding travel behaviour is important for many applications such as studying tourist activity (Edwards, et al., 2009) or the impact of a strike on transportation systems (Tsapakis, et al., 2012). To understand travel behaviour, some standard data collection practices have been in place in order to collect travel data. Among others, travel surveys are one of the most important ways of obtaining critical information needed for transportation planning and decision making. These surveys gather information about the trip-making characteristics such as the mode of transport and trip purpose along with demographic and socioeconomic characteristics of individuals and households. Nevertheless, they are also used to enhance our understanding of travel in relation to the choice, location, and scheduling of daily activities. This enables us to enhance our travel forecasting methods and improve our ability to predict changes in daily travel patterns in response to current social and economic trends and new investments in transportation systems and services. These travel surveys also play a role in evaluating changes in transportation supply and regulation as they occur (Griffiths, et al., 2000).

Travel surveys used to be conducted using different methods such as telephone and face-to-face interviews and computer-based reporting to maintain a diary (Stopher & Greaves, 2007). These have proven to be a burden for participants to use as well as being expensive and time consuming for the data collector (Stopher & Metcalf, 1996). Recently, a new trend emerged as to use GPS devices in conjunction with traditional surveys (Stopher, 2008). Using GPS devices has proven to minimize trip under-reporting through improved survey methods (Bricka & Bhat, 2006). GPS-based surveys were also found useful to record the exact time and destination, as well as to capture underreported trips (Wolf, et al., 2003; Guensler, et al., 2006)). Such surveys are used to produce various analyses such as travel time, modal split, discrete choice and trip distribution models (Ortúzar & Willumsen, 2011). These models include measuring variables such as traffic congestion (Bachman, et al., 2012), properties of mode of transport, length of stages, journey time, purpose of the journeys (DfT, 2011), and so forth.

A further step was to base the diary on GPS devices and subject the respondents to undertake prompted recall surveys in the form of phone or face-to-face interviews. The process of using prompt recall surveys however still proved to be expensive, time consuming and burdensome for both the survey coordinator and the participant. Many research groups tackled this problem by attempting to infer travel information from the GPS data automatically (Liao, et al., 2007; Zheng, et al., 2008; Bolbol, et al., 2012a). Among these types of information are for

¹ Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T, Tsapakis, I & Haworth, J, 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems*, Special Issue: *Advances in Geocomputation*, Volume 36, Issue 6, pp. 526–537.

² Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T, Tsapakis, I and Chow, A, 2012. Sample Size Calculation for Studying Transportation Modes from GPS Data, *Procedia - Social and Behavioural Sciences*, Volume 48, 2012, Pages 3040–3050.

example the transportation mode (e.g. *cycle, walk, bus* and so forth) and the trip purpose (Stopher, 2008). This inference would eventually replace or complete a lot of the feedback required by participants when labelling and tagging their travel diaries.

The classification accuracies achieved by many of these studies vary from as high as 95% (Stopher, et al., 2008a) to as low as 70% (Bohte & Maat, 2009). Nevertheless, due to the diversity of these attempts, different methods have strength-points and limitations within different phases of these studies. As demonstrated in Figure 2.1, some of the limitations influence the **validation** process by which the accuracy level is calculated. On the other hand, other limitations affect the accuracy level achieved due to the efficiency of the **method** used.

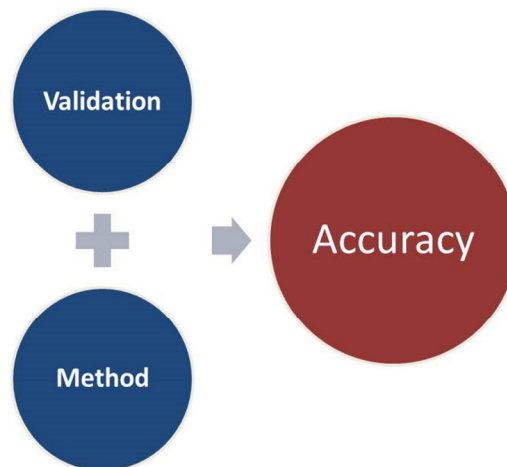


Figure 2.1 Limitation Types affecting the Accuracy achieved by Methods attempting to Infer Transport Mode from GPS Data

Furthermore, as illustrated in Figure 2.2, the limitations that affect the **validation** method can be further subdivided into **data collection** and **data preparation** issues. Contrariwise, the limitations that affect the efficiency of the method used to infer the mode of transport from GPS data can be categorised into three approaches/phases, namely; techniques used for **classification**, techniques used for **segmenting** a GPS track into trips and further into stages, and techniques using **transport network datasets** to directly infer the mode. The following two subsections (2.1.1 and 2.1.2) describe these issues and how they contribute to the research wealth of the methods aiming at detecting the mode of transport and the type of limitations that exist as a consequence.

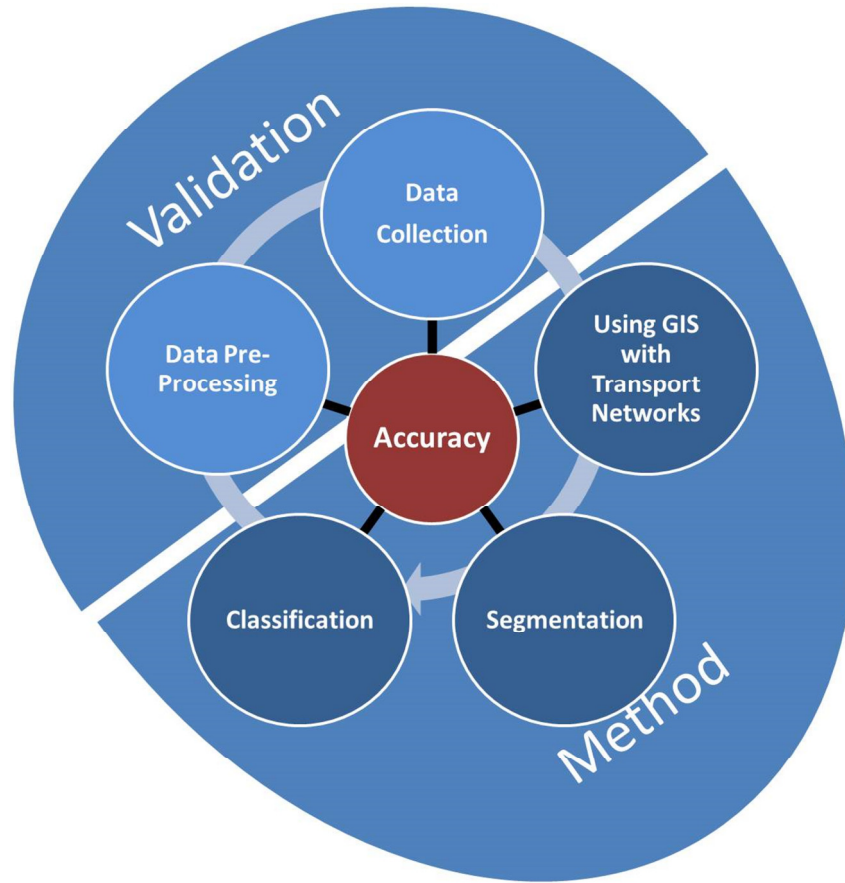


Figure 2.2 Phases affecting Calculated Accuracy of Methods attempting to infer Transport Mode from GPS Data

2.1.1 Effect of Validation Parameters on Accuracy

Different studies attempting to deduce the mode of transport from GPS data assess the efficiency of their algorithms differently. Hence, the classification accuracy calculated by these studies cannot be compared to one another. These studies often choose very different **data collection** parameters such as the device type, the sample spatio-temporal distribution and study duration, the collection rate of the GPS device or the nature of the study area being urban or rural. The top part of Figure 2.3 shows these parameters as part of the data collection planning phase that affects the calculation of the accuracy of the performance of the proposed method.

On the other hand, the lower part of Figure 2.3 shows how the **data pre-processing** phase affects the calculated accuracy. This effect stems from different decisions on how to filter the data based on its quality, the modes of transport that will be considered for the study, the extent to which the trip information will be generalised (e.g. whether only to include most dominant mode for a trip), and the strategy by which the data will be labelled (e.g. whether by participants, and if so how).

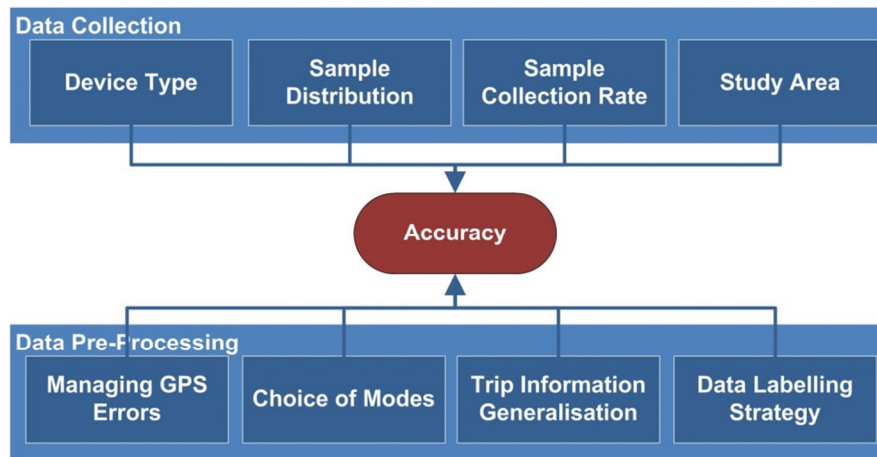


Figure 2.3 Factors affecting Validation of Transport Mode Detection Methods from GPS Data

Therefore, identifying the best values or practices for these parameters would lead to unifying the validation process. This also leads to setting a valid process for calculating the accuracy of methods aiming to classify GPS data into modes of transport. With the validation problems identified, the remaining problems affecting the accuracy are directly related to the method used to identify the mode of transport. This is discussed in the following subsection.

2.1.2 Effect of the Inference Techniques on Accuracy

Previous studies present very different approaches to achieve an accurate classification of GPS data into mode of transport. Some of these attempts either carry out the **classification** taking a pure Machine Learning (ML) approach by using a variable or group of variables, or use a procedural programming approach (Liao, et al., 2007; Wang, et al., 2010; Reddy, et al., 2010; Manzoni, et al., 2010; Stenneth, et al., 2011). Others build up on that by first **segmenting** a GPS track into trips and furthermore into stages, then follow this up by one of the previously mentioned approaches (Stopher, et al., 2008a; Schüssler & Axhausen, 2009; Zheng, et al., 2010). On the other hand, some studies take a pure GIS approach by using **transport networks** to either enhance GPS positioning accuracy by snapping to these networks (Map Matching), or by snapping to the nearest network and hence deducing the mode of transport (Network Matching) (Chung & Shalaby, 2005; Tsui & Shalaby, 2006; Bohte & Maat, 2009; Stenneth, et al., 2011; Gong, et al., 2012).

The impact of these different approaches on the achieved accuracy is very high since they all approach the problem very differently. Each of these attempts achieves high accuracy with respect to certain modes than others, which depends on the **processes** or **sequence of processes** within the inference framework. For example, frameworks that are only network matching-based tend to identify modes that solely use networks better such as the train mode. Another example is studies that base classification on segmentation by clustering for stops often misidentify switches between modes such as getting off a bus then a walk mode since there is no an actual stop made. Another reason why these approaches highly differ in accuracy is due to the techniques that these methods use. These techniques often introduce certain advantages as well as limitations that affect the achieved accuracy. These techniques could be broken into the three previously mentioned approaches, namely; **classification**,

segmentation and using **GIS** with transport network datasets (or as a verification stage). Some studies use either one of these approaches or a combination of two of these processes together. The advantages and limitations introduced within these three categories are summarised as follows in Figure 2.4 and are explained as follows.

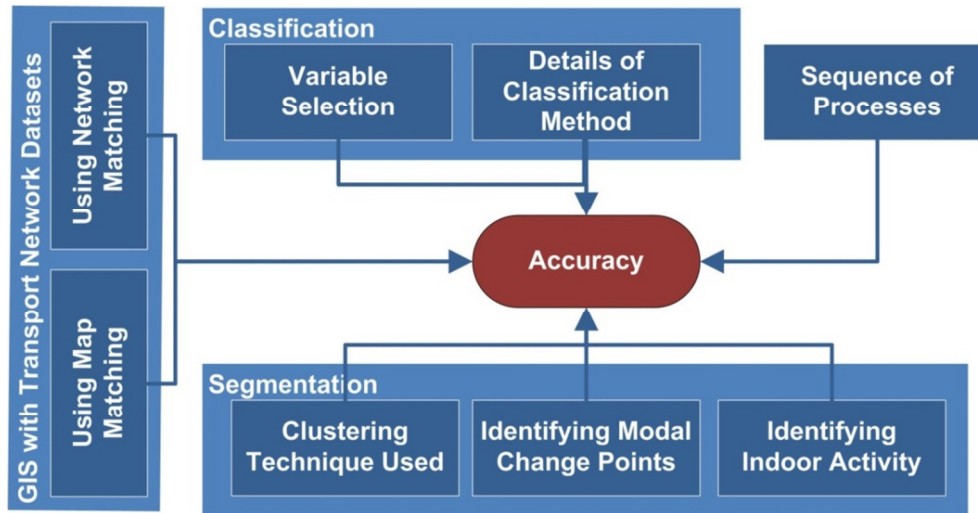


Figure 2.4 Processes affecting Accuracy of Transport Mode Detection Methods from GPS Data

In **classification**-based methods, two main issues need to be addressed in order to ensure that the method is adequately addressing the classification framework. The first issue is the choice of variables extracted from the GPS data that will be used to enter into the classification algorithm. More importantly, if there was any statistical technique used to assess the adequacy of these variables to be used. The second issue is the technique itself used for the classification of GPS data into modes of transport. Since classification techniques possess different properties, the developed classification framework needs to benefit from the chosen technique by applying the technique to the data in a way that takes advantage of these properties. Conversely, the framework also needs to ensure that the limitations of the chosen technique are accounted for.

As for **segmentation**-based methods, the main concern is how the segmentation will be integrated into the designed framework. The segmentation of the track into stops and stages is often used as a complementary process to the classification process. However, some studies only focus on breaking down trips into different segments for travel survey analysis purposes. This kind of studies carry out three main operations, namely; clustering stops in GPS data, identifying modal switch points, and/or identifying indoor activity. These operations will depend mainly on the type of technique used and the limitations that each technique possesses. More importantly, the way that the study defines a stop is key to the performance of that technique.

Finally, in **GIS**-based methods, two kinds of studies should be looked into. The first is the kind of studies that attempts to use transport network data to snap to in order to enhance the accuracy of the GPS fixes as a result of GPS errors. This method is known as map matching, or matching to a transport network. These studies eliminate a big part of the GPS error and

hence increases the accuracy of the variables used for classification which in turn increases the classification's accuracy. The second kind of studies attempts to do the classification by choosing which transport network to snap to in the first place. We coin this method as network matching, since no previous study has assigned any known name for the process. These studies search for the most probable network a stage of GPS fixes belongs to, such as snapping a train trajectory to the train network. As this method might appear very useful in detecting the mode of transport, however, applying it in an inadequate manner might lead to many errors.

2.1.3 Chapter's Content

Therefore, this chapter presents a walkthrough on how these different aspects were handled by the research community and put into perspective. The sections which constitute this chapter are based on the different elements constituting Figure 2.2 where each section discusses one of these five highlighted issues.

Section 2.2 discusses how previous research handles the data collection process describing sampling approaches and how they define the sufficient size and the duration of the sample to be used for testing the inference performance. It also discusses the collection rate (temporal granularity) by which the GPS devices are set to collect data for different studies. It also briefly touches on device types and the study area to be studied.

Section 2.3 highlights data pre-processing challenges such as the method of tagging the collected data and choices made by previous research on what modes to include and the manner by which the trips are tagged. The section also discusses GPS data problems as result of bad GPS accuracy and coverage; and proposes different ways to overcome or avoid them based on previous research.

Section 2.4 describes attempts to classify positional data (such as GPS data) into different modes of transport. The section describes what variables these methods use to classify GPS data based on and why they have chosen them. The section also discusses different classification techniques that are used in these studies and what was the reason behind the usage of such methods. We also highlight the advantages and limitations each method possesses.

Section 2.5 then discusses work done on identifying different breaks in a GPS track. The section describes the different attempts to perform operations like GPS data clustering, identifying modal switch points and indoor activity. Regardless the main aim of these studies, identifying these occurrences can help segment the track by identifying stops and stages of a GPS trip. These studies adopt certain techniques and strategies to identify these instances that carry both advantages and limitations that we highlight in light of the best practices and recommendations. The accuracy of the segmentation and classification in these studies will depend on the how these methods define what is to be a stop in a GPS track. This is also discussed within the section, identifying different assumptions previous studies have made while carrying out the segmentation.

Finally **section 2.6** discusses the adoption of GIS in methods aiming to identify the mode of transport. The section discusses different techniques used in previous literature to use

transport networks to perform map matching to enhance GPS accuracy, and hence, enhancing the mode detection accuracy. The section also describes the work done to perform network matching to transport networks to directly identify the mode of transport from GPS data highlighting the advantages and limitations each method possesses.

2.2 GPS Data Collection

This section describes different parameters that previous studies considered for the data collection process to test the methods they developed to infer the mode of transport from positional data. These parameters include the type of device used to collect data, the sampling strategy, study duration, data collection rate, and study area. Therefore, this section mentions the **devices** used in this type of studies and the **study areas** along with a discussion on their adequacy for testing such developed methods. This section also describes the different sample strategies used to test the developed methods. The discussion sheds some light on the chosen **sizes and durations** of sample testing data with respect to the spatial extent, while discussing the techniques used to calculate such sample sizes. We then discuss how previous research handles choosing the most adequate **rate of collection** (or epoch rate/temporal granularity) by GPS devices.

2.2.1 Travel Survey Definitions

In order to de-construct a GPS track from a mode of transport perspective, some definitions have been standardised to be used for the description of different fragments of the *trip* in travel survey studies. As demonstrated in Figure 2.5, the route between any two consecutive GPS points is called a *segment*. Trips also consist of a number of *stages* (a group of segments). A new stage is defined when there is a change from one mode of transport to another, or where there is a change in vehicle of the same mode (Anderson, et al., 2009). A point separating two stages is called a mode change point (also called mode switch), and the first and last points of a trip are called trip ends (also called origin and destination of a trip). Moreover, any cluster of GPS points within a stage is considered to be a stop.

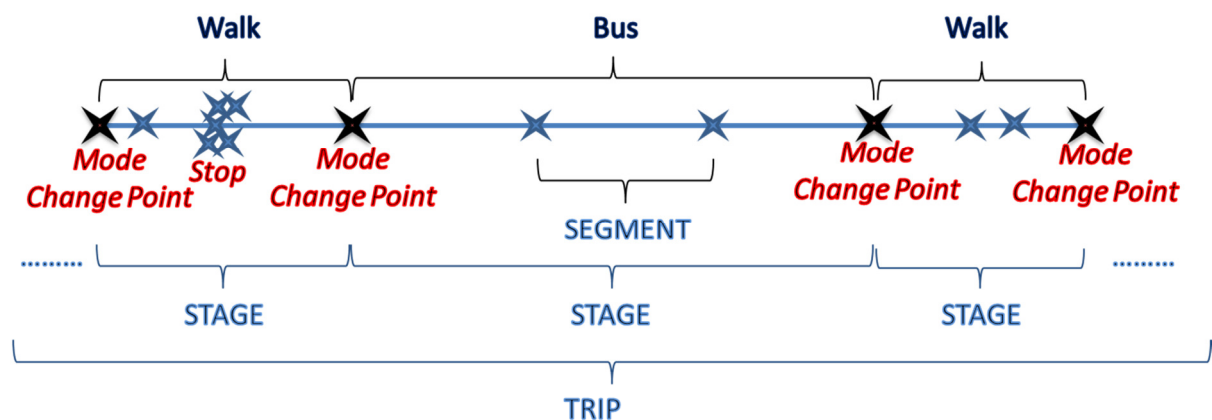


Figure 2.5 Trip Segmentation Elements based on Mode of Transport

2.2.2 Used Device and Study Area

Over the last decade, a plethora of studies have attempted to infer the transportation mode from GPS data collected by travel surveys. Most of these studies have been carried out in complex urban study areas using either: mobile applications on smart phones (Manzoni, et al., 2010; Stenneth, et al., 2011); strictly GPS devices alone (Chung & Shalaby, 2005; Liao, et al., 2007; Stopher, et al., 2008a; Schüssler & Axhausen, 2009; Zheng, et al., 2010) or integrated with other devices, such as accelerometers (Reddy, et al., 2010), or; others through mobile phone call detail records (CDR) (Wang, et al., 2010).

Table 2.1 lists some of devices used by these studies along with the accuracies achieved, sample sizes and durations used for validation, collection rates, study areas and the number of modes considered while validating each method's performance.

| Study | Accuracy (%) | Sample Size (participants) | Duration | Epoch Rate (Sec) | Sensor | Study Area | No. of Modes |
|--------------------------|--------------|----------------------------|----------|------------------|-----------------------------------|------------|--------------|
| (Chung & Shalaby, 2005) | 92 | 60 Trips | - | 1 | GPS | Toronto | 4 |
| (Liao, et al., 2007) | 90 | 4 | 6 Days | | GPS | | 3 |
| (Manzoni, et al., 2010) | 82 | 5 | NA hrs. | 1 | Mobile App | | 7 |
| (Reddy, et al., 2010) | 94 | 16 | 7.5 hrs. | 1 | Mobile App (GPS/accelerometer) | | 5 |
| (Stenneth, et al., 2011) | 93 | 6 | 3 hrs. | 1/30 | Mobile App | | 6 |
| (Stopher, et al., 2008a) | 95 | - | 5 Days | 5 | GPS | Sydney | 4 |
| (Wang, et al., 2010) | - | - | 5 Days | Fine | Mobile Call Records | Boston | 3 |
| (Zheng, et al., 2010) | 76 | 65 | 10months | Fine | Self-Collected | 28 Cities | 4 |

Table 2.1 Summary of Previous Studies' Accuracies, Sample Sizes & Durations

This diversity of positioning sensor devices used for this type of studies is overwhelming in the capabilities that each encompasses. However, the selection of the most adequate device for the study must maintain the threshold between different properties such as the battery life, size, accuracy, coverage, interactivity and real-time capabilities. This could be noted from the selected devices for data collection in these previous studies, and is an important factor in the device selection process.

From the table, we can also note that study areas varied from a city to several cities. However, some studies test the developed algorithm in non-complex areas (Ren & Karimi, 2009). On the other hand, others do not test the algorithm on a large portion of the transport networks of the study city due to the limited amount of the data collected. This is attributed to the small sample size chosen for the validation of the performance of the method.

2.2.3 Sample Spatial Size and Duration for Mode of Transport Inference Validation

Calculating the minimum adequate sample size is an important consideration for the kind of inference models discussed in this research, and depends largely on the variables under investigation in these models that will be used for classifying GPS data into modes of transport (Ortúzar & Willumsen, 2011). For conventional one-day or two-day travel surveys, sample size procedures are well known and widely applied; an example of these standards is the Travel Survey Manual by Cambridge Systematics (1996). The corresponding sample size procedures for GPS-based panel surveys however, are less well developed.

Studies aiming at inferring the mode of transport from GPS data result in collecting a large amount of diverse data to test different approaches. The majority of the studies however did not present the basis of their sampling technique or approach. From Table 2.1, we can note that some of them used as low as 4 participants (Liao, et al., 2007), 60 trips (Chung & Shalaby, 2005) or as many as 4882 participants (Schüssler & Axhausen, 2009) without providing statistical justification for such figures nor sampling justification regarding its spatial distribution. Another issue is the study's duration, where several studies use less than even

one day's worth of data (Reddy, et al., 2010; Manzoni, et al., 2010), whilst others use data of less than a week duration (Liao, et al., 2007; Schüssler & Axhausen, 2009). Note that neither cases account for the weekly seasonal variation which creates a bias towards that specific week of collection. Although other methods using labelled data have achieved accuracies of 90% or more, yet sample sizes, durations or distributions were often not studied to give full accreditation to the results.

One of the few studies tackling this problem is Xu (2010), where it develops a framework to estimate the effective sample size of GPS-based panel surveys in urban travel behaviour studies for a variety of planning purposes. The study attempts to obtain reliable means for key travel behaviour variables such as demographic characteristics and seasonal factors, using data from only 95 households. Stopher et al. (2008a) also attempts to find the best threshold between the minimum sample size and the least sampling period. The study analyses hypothetical and actual multi-day data on person kilometres travelled (PKT), trips, and daily travel time for about 70 persons living in Adelaide, Australia and a second sample of about 500 persons also living in Adelaide. However, to the knowledge of the authors, none of the studies which calculate the sample size for GPS-based travel surveys take into consideration the modal, spatial or temporal granularities of the designed study. This is a clear gap in understanding the spatial properties of the sample required process across different parts of the study area. Another gap also exists in doing these calculations for studies interested in data that is collected in the daytime travel such as that needed for analysis by transport agencies rather than attempting to understand the spatio-temporal extent of the investigated dataset.

2.2.4 Data Collection Rate (Temporal Granularity)

Another issue is the temporal granularity of the GPS data (also called the epoch rate of collection), where most studies use a 1 second collection rate as could be noted from Table 2.1. Although it might seem that the more data the better, however, not only would that create a load on memory and on battery restrictions on current GPS devices or smart phones, but it will also add to the computation cost of any of the used algorithms. This will also impose a daily (if not hourly) burden on the participants to charge their devices and acts as a constant reminder that they have a tracking device. This gives rise to typical participant-related reported problems, such as feeling vulnerable when carrying the device, or influencing their normal behaviour (Anderson, et al., 2009). Furthermore, the epoch rate chosen has a significant impact on the GPS data which results in errors in travel behaviour analysis such as "Cold Starts" (i.e. the device doesn't begin recording at the exact start/end location of a trip) or "Miss-Map Matching" (i.e. problems of lateral movement of position from the GPS trace) (Stopher, 2008).

There are attempts in finding the most suitable epoch rate for monitoring animals (Perotto-Baldivieso, et al., 2008), however, not much research give detailed analysis of the best GPS epoch rate setting for pedestrian activity modelling. Table 2.1 provides a summary of the epoch rates used by different previous research work attempting to deduce the mode of transport from positional data showing the type of sensor used for data collection and the area where the study was conducted. It might be noted that all mentioned studies use the finest grained temporal granularity for collecting their data without any statistical or logical justification.

2.2.5 Summary of Data Collection Issues

Through this section, we have introduced the different common data collection practices of the previous studies aiming at inferring the mode of transport from positional data along with their limitations. The variety of devices used for collecting positional data in previous studies stems from the properties needed by each study's specific requirements; having a long battery life, small size, good accuracy, providing certain coverage, being interactive, or having real-time capabilities. The study areas chosen for testing such algorithms often have limitations such as being very simple or having data not covering many areas of its transport networks.

Sampling issues also include that most studies did not provide an understanding of their coverage and sample characteristics. Many of these studies did not either account for the weekly seasonal variation which creates a bias towards that specific week of collection. Moreover, none of the studies which calculate the sample size and duration for GPS-based travel surveys take into consideration or attempt to understand the modal, spatial or temporal granularities of the designed study. Another sampling problem is defining the GPS data collection rate, where a very fine rate of collection holds limitations. These limitations include overloading the memory and battery power of current GPS devices or smart phones, as well as adding to the computation cost of any of the used algorithms which gives rise to problems like increasing the burden on the participants to charge their devices very frequently and act as a constant reminder that they have a tracking device. Not much research provides detailed analysis of the best GPS epoch rate setting for tracking participants within an urban environment. Moreover, all studies aiming at inferring the mode of transport from this data tend to use the finest grained temporal granularity for collecting their data without any statistical or logical justification.

2.3 GPS Data Pre-Processing

Once an appropriate device, study area and sampling parameters are defined, the following step is to simply collect the data. The data collection phase is a very complicated and erroneous one due to the limitations of GPS technology. Therefore, a very efficient data pre-processing strategy is essential for maximising the benefit of the usage of such data. Data pre-processing is an extremely important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects (Pyle, 1999). Henceforward, the **coverage of GPS technology** is an important problem that research that aims at inferring the mode of transport deals with in different ways. These errors can affect the performance of the algorithm greatly due to the addition of inaccurate data that eventually leads to achieving inaccurate results. These sources of *GPS errors* or *lack of coverage* are presented in this section, illustrating how previous research process and deal with them.

Other decisions are also made regarding **contextual information** that the data is to hold. Among this information is the *choice of which modes of transport* to be considered, the *degree of generalisation of trip information*, and the *labelling strategy* of the GPS tracks into modes of transport. These decisions heavily affect the validation process since the inference method developed will test its results against the contextual information added to the track. The calculated accuracy will then partially depend on the accuracy of the provided information. Therefore, we discuss later in this section strategies followed in previous research regarding these decisions, highlighting the best practices and limitations that lead to an inaccurate validation process.

2.3.1 Dealing with GPS Limitations

Around 86,000 data points could be collected for a single day for any individual, given that positional data is collected at a 1 second epoch rate. According to some studies, an average person travels around 70 minutes a day (Stopher, 2008). With this great amount of positional data, GPS errors could bear a massive effect on the data analysis due to the positional inaccuracies. This is usually due to GPS systematic errors such as (Jun, et al., 2006):

- Having too few satellites in view
- Having the satellites aligned in a bad geometry (High HDOP)
- Multi-path signal reflection
- Ionospheric effects
- Clock or receiver issues
- Signal blocking

In this research, GPS data is used to detect the mode of transport of an individual from trips he/she does in his/her everyday life. There are many issues with processing GPS data for this purpose, mainly is collecting data with **low positional accuracy** due to the conventional GPS sources of error mentioned above. This issue affects variables such as the calculated speed and acceleration as well as processes such as the map matching process and identifying stops (Stopher, 2008). Here we aim to shed some light on these problems and possibly some solutions that were proposed by previous research.

Another issue is that there is too little fixes to understand the trajectory which is caused by **loss of signal** due to GPS coverage constraints (Hinch, 2007). Therefore, we also describe how previous literature overcomes this problem. We also show that some issues bear some advantages such as loss of signal where sometimes the loss of information is in itself useful information.

2.3.1.1 Wandering Errors & Urban Canyoning

GPS errors decrease the accuracy of the produced data to be used for analysis, which in turn decreases the accuracy of the analysis itself. Even a stationary GPS device will record apparent movement as result of an imprecision of position fix which appears as wandering usually over distances up to 30m or more. According to Stopher (2008), this is called the “**wandering effect**” which could be defined as an excessive position wandering from a stationary device. The source of error that causes wandering and is most troublesome for all means of transport is the multi-path problem, which is typical to appear in urban canyons. This happens as a result of reflection of GPS signals off buildings, walls and surfaces (Wolf, et al., 2003). For travel analysis purposes, this is a major disadvantage and sometimes worthless since neither the route nor stops in the area can be detected (De Jong & Mensonides, 2003).

This effect could be eliminated by setting strict thresholds on position changes, which will still have limitations like if the device is actually moving but had no much satellite coverage. This process could be considered as a filtering phase of the systematic errors. Most GPS devices have built-in firmware that interpolates position after the **number of satellites** decreases below four. Therefore, the first step to overcome this is to only use fixes with no less than 4 satellites into the dataset and removing any other fixes with fewer satellites (Stopher, 2008).

Similarly, a high value of the **Horizontal dilution of precision (HDOP)** indicates that the satellites are nearly in a straight line or are dispersed around the horizon, either of which gives an unreliable position (Hinch, 2007). Stopher et al. (2005) considers a value of 5 to be the maximum, and any value greater than that is unreliable and hence is removed as an invalid fix. This could be considered as a second step in the filtering process.

Schüessler and Axhausen (2009) propose some other criteria that could reliably detect erroneous GPS points while omitting true ones. One of these is the **altitude** value, which could be set to not exceed or decrease beyond a range between thresholds according to the topographic nature of the area under study. This could be considered as a further step that detects outliers without having to study the horizontal motion between points yet.

Yet other indications for erroneous GPS points are **unrealistic speed** and **acceleration** values. Schüessler and Axhausen (2009) detect points with speed 50 m/s or accelerations above 10 m/s² to be erroneous. However, this could depend massively on the nature of the transport network within the study area. Therefore, a possibility is to apply different thresholds according to different transport networks and congestion zones.

A further step to ensure the logic and decreasing the effect of GPS errors is applying a **smoothing** technique to speed or acceleration values. This potentially could detect and

remove outliers from the calculated speed and acceleration values. Jun et al. (2006) presents a smoothing technique that applies a modified Kalman filter and compares it to applying a least square spline and a kernel-based method. On the other hand, Schuessler and Axhausen (2009) apply a Gauss kernel smoothing technique to remove speed and acceleration outliers from the collected data.

Another way to remove the wandering effect off stationary locations is to apply some sort of a **spatial clustering** technique. Spatial clustering of GPS data detects the centroid of any stop while also removing horizontal positioning errors occurring at any stops in the track. Subsection 2.5.1 of this chapter describes in detail previous attempts to perform GPS spatial clustering as part of the process of identifying stops.

2.3.1.2 Signal Loss

Another significant problem in keeping track of a human trajectory in an urban area is signal blocking or loss of signal (Jun, et al., 2006). It could lead to missing entire trips from a GPS track, which could affect travel survey analysis results, and sometimes bias them. One of the reasons of this blockage is sometimes referred to **urban canyons**, where the entrance of participants into areas of tall buildings could lead to a serious block of the satellites visible in the sky (Stopher, et al., 2005). Another reason is the usage of certain **types of transport** such as cycling, walking and travelling by car, have reported to provide good GPS coverage while it varies considerably for public transport journeys, depending on the proximity of the person to the nearest window within the transportation carriage (Draijer, et al., 2000; De Jong & Mensonides, 2003). A third reason is due to the entrance of the participant into a building or generally **indoors**, or travelling through a **tunnel** or **underground**. A last reason could be due to the **form of carrying the GPS device**.

Hence, signal loss could be categorised into partial and total blockages. **Partial blockage** occurs when less than 4 satellites are captured in the fix and that will be due to an urban canyon or usage of public transport. Conversely, a **total blockage** occurs when no satellites are captured at all and happens upon entering indoors, underground or into a tunnel. Partial blockage is usually a burden and research tends to overcome such inconvenience, where total blockage can be used to identify indoor activity and underground travel.

One solution to avoid partial blockage is to advice participants to sit near windows whilst using public transport to avoid partial blockage. This however could affect the participant's behaviour which might create a bias in his/her actions than the norm.

In order to overcome partial blockage, Wood and Mace (2001) propose combining GPS with a dead reckoning-based positioning system. Dead reckoning is the process of determining position by projecting heading and speed from a past known position. These systems use some type of gyro to calculate speed and heading to determine a differential position. Unfortunately, such equipment is usually expensive and less portable to be used by a large number of participants.

Nowadays, with the rise of smart phones, engines such as Skyhook Core Engine which is a software-only location system could determine the device location with 10 to 20 meter

accuracy (Gallagher, et al., 2009). The Core Engine collects raw data from Wi-Fi access points, GPS satellites and cell towers with advanced hybrid positioning algorithms. So whenever the GPS module gives no solution, the device uses the other two modules to calculate its position. The only disadvantage while using positioning from smart phones is the battery's low life time.

On the other hand, to benefit from total blockage, Stopher et al. (2005) attempts to reason if the signal loss period contains any stops and hence a split in the trip. Therefore, it develops a stepwise algorithm to identify the existence of stops within the signal loss period, and it also detects the most probable route taken after identifying whether the participant was using a car or public transport by reasoning using a GIS transport network. The judgement of stops existence is based on a comparison of the estimated speed with the actual speed taken to enter and exit the urban canyon. However, none of the previous studies to the author's knowledge used transport networks to differentiate underground from car mode of transport.

2.3.2 Choice of Modes and Degree of Trip Information Generalisation

Another decision need to be made by the data processor regarding the level of detail of information the data will have. Such information that needs to be decided upon includes the modes of transport the data will be labelled with and the level of generalisation the trip information will be decided upon.

2.3.2.1 Selected Modes of Transport

The majority of previous mode inference studies only consider a limited number of modes of transport. As shown in Table 2.1, some studies consider as few as 3 modes (Liao, et al., 2007; Wang, et al., 2010), while most studies exclude the train and underground modes. Others generalise the motorised modes together (Reddy, et al., 2010) grouping bus and car modes vehicle modes. This is considered to be a major limitation as not only does this dismiss identifying more information, but it also leads to a high confusion as a result of the different nature of these modes.

2.3.2.2 Trip Information Generalisation

Some of the current studies have other data processing decision limitations, such as using single-purpose trips (Stopher, et al., 2008a; Manzoni, et al., 2010) that influence the results of any inference of any method used, restricting the outcome to one or two modes. Other studies assume that each trip consists of only one travel mode which is the most dominant mode for each trip (Stopher, et al., 2008a). This leads to generalising trip information, and hence, decreases the accuracy of the process of learning and validation. An example of that is having a trip with a train and bus stages, and a decision on counting the trip as once stage of train since it covers most the travelled distance for the trip while ignoring the bus and walk stages.

A long standing argument would be what to count as a stop (Stopher, et al., 2005). Stops can occur before or at an intersection or an extended traffic stop, or within any network at bus stops, train stops, etc. However, it has to be clearly defined as to what is to be counted as a stop and in which scenarios would it be a stop and when would it not be one. For example,

stopping at a cash machine (outdoors) for a short while to withdraw cash can be considered as a stop or ignored. Therefore, while adding segmentation metadata to a participant's track, these fuzzy decisions should be made clear to the editor (being the participant or otherwise). Only by standardising how participants tag their tracks with respect to stops will we be able to accurately assess the efficiency and accuracy of identifying these stops and ends.

According to Stopher (2008), Stops made using in-vehicle or portable devices could be purely for traffic control purposes, to pick up/drop someone, to order/pick up items at a drive-through facility, to stop to drop a letter, or to buy a coffee/newspaper. Moreover, other longer stops depend largely on the main purpose of the trip. For example, a "commuting to work" purpose will mean that the destination is the workplace of the participant, while a "leisure" purpose will mean that the gym for example is the destination of the trip. Moreover, another problem is that most previous studies assign different trip purposes. In London, TfL (2009b) for example lists purposes as: work, education, shopping, personal business, leisure and other. In Canada however, R.A. Malatest & Associates Ltd (2012) adds recreational, social, dining, pick up and drop off as trip purpose categories. It could be noted that there is a clear conceptual overlap between the two definitions between stops and trip ends. Therefore, an important question to ask is "What should be recorded as separate activities from the travel, so that a consistent definition of trip purposes can be achieved?" And the answer to that would depend on the amount of trip detail required. Hence, a research that aims at detecting these segmentation features needs to first clearly define what is to be considered as a stop. That will lead to a clearer and more accurate segmentation, and will minimise problems such as confusing spending 5 minutes indoors in an underground station with reaching an indoor destination.

2.3.3 Data Logging Strategy

Along with GPS data, metadata such as the purpose of the trip and travel modes (walk, bike, car etc.) is usually captured by participants using travel diaries during or after the journey using telephone surveys or face-to-face interviews (Stopher & Metcalf, 1996). This metadata is used to validate the performance of the inference algorithms. However, some of the previous methods did not seem to base their validation on metadata labelled by the participants. For example, Schüssler and Axhausen (2009) use consensus data from previous years of the same city to evaluate the classification results, while Wang et al. (2010) use Google Maps to verify the results by comparing them to the proposed modes by Google for the corresponding trip travel times. The reason why these previous studies avoid collecting metadata is because it has been proven to be a burden on users, achieved low response rates (Bricka & Bhat, 2006), hold many privacy issues (Wolf, et al., 2003), and many other users do not recall exact details, as well as being slow, expensive and time consuming for the data collector (Stopher & Metcalf, 1996).

As the whole world is going online, many long processes that used to be carried out very difficultly (sometimes manually) need to be brought to a digital, more dynamic and interactive environment via the web. Among these is the collection of travel information. The use of an Internet recall survey is more complicated, but offers useful opportunities for further processing of the data during the validation by the participants. Among others, the Internet is

a medium that enables us to present derived data in interactive maps and tables. A system based on web-participant-based reporting would facilitate the processing of GPS data into trip information quickly and cheaply, since no manual data entry is required. Stopher and Collins (2005) conducted a pilot study investigating the improvement of data collection using GPS data loggers in vehicles by tagging trip information using an Internet recall survey. They created maps depicting routes, origins and destinations per respondent for each day. The respondents were able to indicate missed trips and destinations on the map, which were then processed together.

Doherty et al. (2006) anticipates that respondents cannot handle interactive maps, and therefore methods that require the interpretation of maps, such as that introduced by Stopher and Collins (2005), are not very user friendly. Therefore, Doherty et al. (2006) proposes that GPS data are first split into trips, then activities and missing sections are determined via algorithms and subsequently start and end times, modes of transport used and activity locations are determined with the use of transport networks. The results are then presented to the respondents in tables in an Internet questionnaire. The respondents are asked to check the details and to add information relating to the number of travelling companions, amongst other things. It is possible to view the trips on a map in the application should respondents wish to do so, but any adaptations to the data by the respondents are made in the tables.

To the knowledge of the author, no complete application was found that serves this specific purpose. However, there are a number of similar applications that have different purposes and accordingly have different functionalities to serve that purpose. The MapmyFitness Inc (2013) application enables users to upload the GPS tracks and to state which type of sport they were undertaking. It returns analysis on their workout along with a representation of their tracks on a Google Maps background. Innersource (2013) also enables users to view their GPS tracks, routes and way points on different maps background for sharing and social interaction. Another similar interface is GlobalMotion Media Inc (2013) for sharing travelling experiences around the world on the web. Very few applications provide the functionality of editing GPS tracks online from a map interface such as Schneider (2013). Also, some desktop applications enable users to perform edit functionality (Dice Holdings Inc, 2013), but yet not having the functionality of adding any attributes to the data.

Bohte and Bhat (2009) is the only study that develops an interface that enables participants to browse their tracks chronically on a map interface. The trips are displayed as a table that lists the originating location, the departure time, the mode used, the arrival time and the trip purpose. The interface then enables the users to adjust any wrongly inferred trip metadata or approve of the information reviewed. The interface also allows users to split or merge trips or to move the location of a trip destination and adding whole trips. However, this study bears some disadvantages such as the usage of local map authority's basemaps that are not familiar to the users, and not taking advantage of recent map visualisations such as the Google Streetview service. Another limitation is interoperability, where users are not allowed to upload or manage any of their own data, which is advantageous for facilitating the process for the users yet disadvantageous for limiting the interaction by not providing the option for manual processes. Moreover, a potential of using incentives such as sharing tracks using social

media, providing travel statistics, or CO₂ emissions could have been used to increase the response rate and reporting accuracy.

2.3.4 Summary of Data Pre-Processing Issues

As discussed in the beginning of this section, we explain the guidelines and practices previous research used to dealing with and benefiting from GPS limitations by filtering and processing GPS data. We demonstrated that there are two cases of implications of GPS technology limitations, namely; low positional accuracy and signal loss. Previous research has dealt with low positional accuracy overcome problems such as wandering errors & urban canyons by filtering data by dismissing fixes with less than 4 satellites, HDOP more than a value of 5, unrealistic altitude value, speeds above 50 m/s or accelerations above 10 m/s² according to different congestion zones or transport networks. Other studies apply smoothing techniques for eliminating urban canyon effects or spatial clustering for wandering effects. Spatial clustering is also used in segmentation by identifying stops, and is further discussed in section 2.5 of this chapter as it is dedicated to reviewing previous attempts to segment GPS tracks into trips and stages.

As for signal loss, it is categorised into partial and total signal blockage. Previous research has attempted to overcome partial blockage by integrating the GPS device with other devices such as gyros, Wi-Fi, or GPRS technologies. The disadvantage of these hybrid solutions is that such equipment is usually expensive, less portable, or consumes high battery power. On the other hand, total blockage has been treated by using transport networks to find missing segments of a GPS track. Total blockage has also been used to detect indoor occurrences, which is described in detail in section 2.5 as part of segmentation approaches of tracks into stages. Total blockage could be also beneficial for identifying the certain modes such as underground travel. Work attempting to identify underground travel is described in detail in section 2.6 as part of describing attempts to use transport networks for inferring the mode of transport.

The validating and testing of algorithms that aim to infer the mode of transport are only as accurate as the information provided with the data. This information includes the types of mode of transport that are to be investigated, the generalisation level of the trip information, and the strategy used to report the mode of transport by the participants. As discussed in this section, some decisions made regarding this information hold many limitations. Among these limitations is considering a limited number of modes of transport for the developed method. This makes these methods less robust to identify ignored modes. Another limitation is that some studies generalise the modes used in a trip to only the most dominant mode within this trip. This decreases the accuracy of the learning and validation process as a result of having a mode being denoted by mixed modes. A clear standardisation of stop definition whilst tagging a track will lead to accurately assessing the efficiency and accuracy of identifying these stops when validating the attained results.

The metadata, such as the mode of transport, used to validate the results of the inference method is usually added by the participants reporting the details of their GPS tracks. Some studies do not use participant-reporting to add this metadata to the GPS data track and instead depend on assumptions that may bear a large amount of inaccuracy. These studies do not collect participant feedback due to reasons such as the being burdensome and bearing

privacy issues for the participant bears, as well as being slow and expensive for the data collector. Instead, some web-applications exist that enable the participants to view their tracks and edit them, yet still bearing several limitations such as not being able to add data metadata. Only one study enables the participants to fully check and edit their tracks. However, the study holds other limitations such as presenting unfamiliar basemaps, not using useful web-mapping products useful for better recall such as Google Streetview, and enabling the user to only use the data the data collector provides limiting the interaction of the participants with the data.

2.4 Mode of Transport Classification Techniques

Studies aiming at inferring the mode of transport could be divided into procedural and Machine Learning (ML) approaches. Procedural approaches attempt mainly to make inferences based on logical assumptions, such as how a typical person would travel (Stopher, et al., 2008a). Other assumptions include the surrounding environment, such as the nearest transport networks (Chung & Shalaby, 2005) or the temporal-related assumptions of activities, such as people are more likely to have no activity after mid-night (Liao, et al., 2007). On the other hand, ML approaches attempt to do the inference based on learning from existing data, possibly combined with similar logical assumptions. Examples of these studies use Decision Trees (Zheng, et al., 2010; Reddy, et al., 2010; Manzoni, et al., 2010), Bayesian Networks (Stenneth, et al., 2011), Fuzzy Logic (Schüssler & Axhausen, 2009), Hierarchical Conditional Random Fields (Liao, et al., 2007), and Support Vector Machines (SVM) (Zheng, et al., 2008). Previous literature suggests that ML methods are more plausible due to their flexibility and their independence from any assumptions on how a human trajectory (or an individual) should act (Zheng, et al., 2008). This is a major advantage since people are different and only a few and very generic rules can be imposed to dictate how an individual acts and moves.

This section describes the details of these previous methods attempting to infer the mode of transport from GPS data. These methods are described in terms of different issues. Many limitations arise in every method in relation to certain issues. Hence, this section mainly highlights the method-related limitations in each of these methods. One very important issue is selecting the variable that best differentiates modes from one another. Subsection 2.4.1 discusses this issue while demonstrating the different variables that previous research use. The other method-related limitations are discussed in subsection 2.4.2 in detail. These method-related issues include the usage of a limited number of transportation modes in the learning, the high dependence on segmentation into transportation modes, and high reliance on temporal information.

2.4.1 Variable Selection

Generally in a classification problem, the variable that is to be predicted is known as the dependent variable (mode of transport in our case) because its value depends upon, or is decided by, the values of all the other attributes (Mitchell, 1997). The other attributes that help to predict the value of the dependent variable, are known as the independent variables (IVs) in the dataset. The less correlated (or statistically dependent) the IVs are the more the outcome of the classification is inclined to be biased.

In the case of this study, the IVs must be chosen to be used in classifying GPS tracks into modes of transport. To achieve the highest classification accuracy, these IVs must discriminate between the different modes of transport very efficiently. Most studies use variables such as length, speed, acceleration, maximum or median of speed or acceleration through a stage (Schüssler & Axhausen, 2009; Zheng, et al., 2008), either together or individually for classification without providing a statistical basis for the choice. This is a major limitation since none of the studies carry out any statistical evaluation for choosing these IVs ensuring that they are the most efficient IVs, or group of IVs, to be used to achieve the highest

classification accuracy. The correlation of the chosen IVs in these studies was neither accounted for.

2.4.2 Classification Limitations

The range of the methods used to infer the mode of transport from GPS data has extended from logical procedural to Machine Learning (ML) approaches in order to resolve a classification problem.

Stopher et al. (2008) uses a process of elimination of different modes at different phases of the algorithm. As simple as this procedural method is, yet it does not account for properties within the data like ML techniques. These properties include studying the sequence of a string of GPS segments together as one continuous chain of movements.

Schüssler and Axhausen (2009) developed an open source fuzzy logic engine using the median of speed, the ninety-fifth percentile of the speed and the acceleration distributions as fuzzy variables. Fuzzy logic gives computers the ability to emulate human reasoning and solve certain types of problems efficiently. The word "fuzzy" refers to the imprecise type of logic used to manage real-world tasks such as the problem in this research. However, computers that use fuzzy logic do not have the ability to learn and adapt after solving a problem as some expert systems can.

Several studies employ decision trees to perform this classification, either alone or integrated with other techniques, such as Hidden Markov Models (HMM) (Zheng, et al., 2008; Reddy, et al., 2010; Manzoni, et al., 2010; Stenneth, et al., 2011). Decision trees are advantageous in that they require no normalisation and hence require very little data preparation, as well as being robust and Performs well with large data in a short time. However, decision trees can be extremely sensitive to small perturbations in the data where a slight change can result in a drastically different tree. They also can easily over-fit, which can be negated by validation methods and pruning even if it were a grey area. Moreover, since decision trees are not as smooth as other ML methods, they can have problems carrying out an out-of-sample prediction (Mitchell, 1997). Furthermore, for data including categorical variables with different numbers of levels, information gain in decision trees is biased in favour of those attributes with more levels.

Zheng et al. (2008) compares between different ML methods to infer the mode of transport from GPS data. Among these method is Support Vector Machines (SVM), and deduces that it does not achieve as high accuracy as the rest of the techniques. Although SVM's limitation is speed and size (Burges, 1998) both in training and testing, yet it bears many advantages. Among these advantages is that it is easily trained and requires no feature extraction, and hence, it efficiently learns from the data structure (Shawe-Taylor & Cristianini, 2000). Zheng et al. (2008) also uses various variables together such as Length, mean velocity, expectation of velocity, covariance of velocity, top three velocities and top three acceleration between GPS fixes; yet not accounting for the track as a continuous chain of movements.

A couple of studies also use temporal information for mode inference. Liao, Fox and Kautz (2007) use the time of day to use in a probability model building assumptions about the participant's context. While this might be a useful technique to identify different activities, it

might not be applicable to participants that have abnormal working hours for example. Stenneth et al. (2011), on the other hand, depends on live bus and train times information to make some inferences too, which would require a continuous input of such information for any period of time.

2.4.3 Summary of Classification Technique limitations

This section discussed the approach different methods have taken in order to carry out the classification of the GPS tracks into modes of transport. A major limitation all methods have is the lack of strategy for the choice of the IV to be used to differentiate between different modes of transport. These studies carry out no statistical evaluation whatsoever for choosing these IVs ensuring that they are the most efficient IVs, or group of IVs, to be used to achieve the highest classification accuracy. The correlation of the chosen IVs in these studies was neither accounted for.

The impact of the different techniques used for classification on the achieved accuracy is very high due to the difference in the nature of the used methods. Most **procedural** methods comprise problems such as not considering the sequence of GPS segment strings together as one continuous chain of movements. This results in losing the pattern of movement of a trajectory which is considered a major loss of information.

On the other hand, **ML methods** have different problems depending on the technique used. Methods using *fuzzy logic* do not have the ability to learn and adapt after solving a problem as some expert systems can. *Decision trees* can be extremely sensitive to small perturbations in the data which results in over-fitting. This is not very useful for detecting the mode of transport since if the method were to use speed there would be a high confusion due to this sensitivity. Moreover, since decision trees are not as smooth as other ML methods, they can have problems carrying out an out-of-sample prediction. Moreover, for data including categorical variables with different numbers of levels, information gain in decision trees is biased in favour of those attributes with more levels.

Another common problem with both procedural and ML methods is the usage of temporal information. While this might be a useful technique to identify different activities, it might not be applicable to participants that have for example abnormal working hours. Jiang et al. (2009) has shown that human mobility patterns are mainly attributed to the underlying street network rather than the goal-directed nature of human movement which has a little effect on the overall traffic distribution. This therefore stresses the need to adopt a Machine Learning method that deeply understands human movement patterns rather than attempting to depend on repetitive trip information over cyclic temporal periods. Although *Support Vector Machines (SVM)*'s limitation is mainly the computational speed both in training and testing, yet it can easily be trained and requires no feature extraction, and hence, it efficiently learns from the data structure. This is specifically useful since it can understand the different pattern by which a trajectory moves in different transport networks.

2.5 Identifying Stops and Mode Switches

This section introduces work done on identifying breaks in a GPS track. These breaks in the track can either be a random stop in the middle of a stage or a switch in mode separating two stages. Therefore, this section discusses different methods aiming at segmenting a GPS track into stages and stops by respectively identifying mode switches and stops within stages. In order to identify any of these two events, this section demonstrates that variations of three methods were mainly used which include clustering, identifying switches and identifying indoor occurrences. This is briefly illustrated in Figure 2.6 where each of these three methods can contribute to identifying points that can be interpreted as one of the two track segmentation events, namely; stops or mode switches (or both).

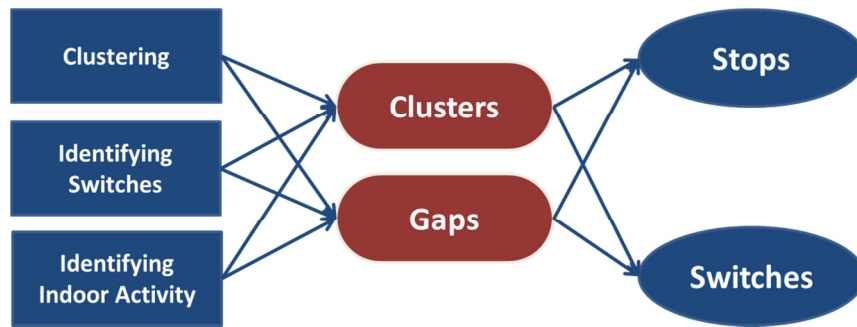


Figure 2.6 Segmentation Processes leading to identifying different Breaks

2.5.1 Identifying Clusters

As mentioned in section 2.3.1.1, the wandering effect often occurs as result of the GPS inaccuracies, which can be defined as an apparent movement of a stationary GPS device might occur (Stopher, 2008). As a result many stops in a GPS track might be appear as a cluster of points. These stops could be either ordinary stops within a GPS stage (e.g. a car stopping at traffic lights), a switch in mode (e.g. walking to a bus stop, waiting for the bus, then taking the bus, arriving at destination). This is demonstrated in Figure 2.6 showing that a cluster can be either one of these two possibilities (stop or mode switch). A stop can also be interpreted as a mode switch depending on the strategy used to define each term.

Many attempts were carried out in order to perform this clustering varying from basic attempts to very sophisticated algorithms. While Schönfelder and Samaga (2003) use in-vehicle GPS data to define trip purposes, they identify stationary clusters whenever the vehicle stops. They attempt to find the centroid of the cluster by searching within a radius of 200 m from each respective point within the cluster for the rest of the points within the cluster (Figure 2.7). The fixes which have most neighbours and the smallest average distance to the all other considered points is classified as the cluster's centre. This, while being simple and computationally low costly, does not account for slow speed trajectories such as walks since it only considers vehicle trajectories. This method does not either account for high density transport networks that consist of many small links close to one another.

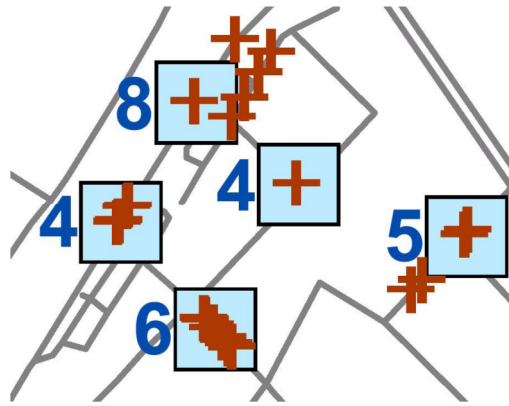


Figure 2.7 Clustering of observed stops in trips (crosses) to unique activity locations (boxes) (Schönfelder & Samaga, 2003)

For wearable portable devices, Stopher et al. (2005) develops a very intuitional algorithm to identify stationary clusters. Where if a group of records are within 30 m of each other they are grouped into cluster. The rationale for using 30 m stems from previous analysis of data records of stationary devices. It was found that the average GPS accuracy was around 10 m and therefore, used three times the standard deviation (30 m), as the critical range to determine when the GPS device starts to move again provides 97% confidence. No level of accuracy is mentioned however on the correctly identified stops. The method still does not account for dense networks or slow speed modes.

Schüessler and Axhausen (2009) build up on the work of Stopher et al. (2005) by putting further constraints on the criteria of selecting a cluster. They define a “point density” by being determined by counting how many points are positioned within a 15 meters radius around it. Then if the sequence of points with a density higher than 15 lasts for at least 10 points or 300 seconds, a stop or stationary cluster is identified. They also apply the “zero speed” criterion adapted from Schönfelder et al. (2006) and Tsui and Shalaby (2006), where if the speed is less than 0.01 m/s for at least 120 seconds, a stop is identified. The study, however, had no actual labelled data to validate against, and hence, they used consensus data of the same period of time using the identified percentage of stops as basis of the validation. The method also does not reason for time differences between points, hence only accounting the spatial while ignoring the spatiotemporal domain.

Ashbrook and Starner (2003) tackle the problem with a more sophisticated approach. They use a variant of the K-means clustering algorithm to identify clusters and their centroids (Figure 2.8). The basic idea is to take each point and treat it as a centre with an assigned radius. All the points within this radius are marked and the mean of these points is found and is used as a new centre. The process continues with the same radius but around the new centre until the mean stops changing, the points within the radius are removed and are replaced with a new “location” with a new “location ID”. The algorithm moves on to the rest of the track until all locations are identified. Different radii were used and the most appropriate radius found was around 300 m. The only validation that was stated however is that identified clusters had a mean distance of 48.4 m to the nearest building, as the algorithm being part of a study aimed at determining users’ significant locations. Also due to this fact, the

spatiotemporal domain was not accounted for while performing the clustering. The study was also conducted on a single user which creates a chance for bias.

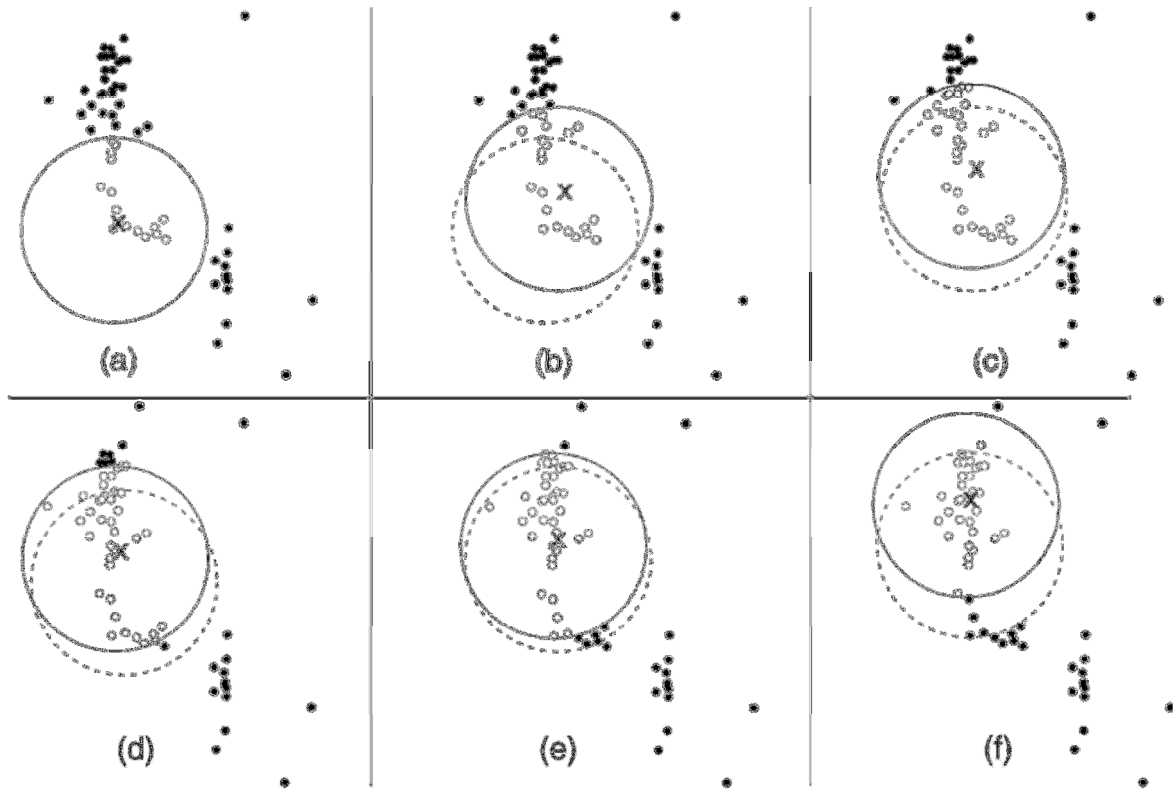


Figure 2.8 Illustration of a location clustering algorithm (Ashbrook & Starner, 2003)

The X denotes the centre of the cluster. The white dots are the points within the cluster, and the dotted line shows the location of the cluster in the previous step. At step e, the mean has stopped moving, so all of the white points will be part of this location

2.5.2 Identifying Change Points

Several studies aiming at inferring the mode of transport from GPS data tend to search for change (inflection) points where the mode appears to change. These studies tend to search for these by searching for change in speeds or by separating them by walking segments. These studies also present a variety of techniques to reason about the switches between one mode and another. This subsection describes methods that attempt to carry out this type of segmentation along with the techniques used to understand transition between one mode and another.

2.5.2.1 Identifying Mode Switches

There are two types of approaches in terms of stage segmentation. One type ignores the segmentation on the whole, while the other performs the segmentation then bases the classification phase on the segmentation outcome (Zheng, et al., 2008; Schüssler & Axhausen, 2009). The second type works on segmenting the track into different stages, each stage with a different mode, then classifying each stage into a mode of transport based on the chosen IVs such as speed. The limitation that these methods present is that the overall classification

accuracy is heavily based on the segmentation accuracy. Another disadvantage of this approach is that if a stage is broken into two or more stages, the classification of these derivatives might differ greatly to the classification of the big stage as a whole. This is due to the changes in the overall movement pattern that a certain mode might have in different spatial environments.

Some of these methods identify the change points using the previously mentioned clustering techniques. Other studies use sudden change in speed or acceleration to identify these points (Stopher, et al., 2008a). However, this can produce many errors such as in situations where a vehicle stops at traffic lights or suddenly approaches a traffic jam. Other studies attempt to first classify the slowest speed movements as walk stages and assumes that between every two walk stages lays a non-walk stage to be classified into a different mode (Schüssler & Axhausen, 2009). Nevertheless, this ignores instances where certain modes move at abnormal speeds such as in traffic jams, smaller cities, rural environments or residential neighbourhoods.

2.5.2.2 Modal Transition

Most studies assume that any two stages are always separated by a *walk* stage (Schüssler & Axhausen, 2009). This, while true for most cases, might fail in cases of *cycling* or driving the *car* out of a *train* station's car park for example. Liao et al. (2007) propose a transitional compatibility between different activities where it simply defines which modes can follow one another. Another useful method to account for this is to use a transition matrix to verify the mode switch between consecutive stages according to a probability matrix of such switches (Zheng, et al., 2008). The study proposes a useful transitional probability idealisation between modes that helps identifying the change points from one mode to another effectively with an accuracy of (89%). It uses a transition matrix between different modes that states probabilities of switching from one mode to another Table 2.2. The study on a whole does not produce high accuracy (65%). However, it only considers 4 modes for the study and performs the segmentation before mode classification, which hugely influences the accuracy of the overall classification.

| Travel modes | Walk | Car | Bus | Bike |
|--------------|-------|-------|-------|-------|
| Walk | | 53.4% | 32.8% | 13.8% |
| Car | 95.4% | | 2.8% | 1.8% |
| Bus | 95.2% | 3.2% | | 1.6% |
| Bike | 98.3% | 1.7% | 0% | |

Table 2.2 Transition Matrix of Modes of Transport (Zheng, et al., 2008)

2.5.3 Identifying Static Indoor Activity

Another approach to segment a trip into smaller fragments is identifying indoor activity. Indoor activity includes all loss of signal occurrences (described in subsection 2.3.1.2). Indoor activity can be divided into *movement* and *static* events. Indoor movement events include movement using underground transport, driving in tunnels, walking in tunnelled areas, or any general loss of signal while travelling. This is disused in section 2.6.2 as part of describing work using GIS data to identify these events. On the other hand, static indoor activity usually reflects being at a destination (or significant place) such as home, workplace, etc. (Stopher, et

al., 2008a; Liao, et al., 2007). This subsection describes methods used to identify where exactly these locations are, and then describes methods used to identify these occurrences.

2.5.3.1 Identifying the Location of Indoor Activity

Determining locations of indoor occurrences in a GPS track could be a difficult process mainly due to the Cold Start effect (Stopher, 2008). The Cold Start effect arises from the situation that the device does not begin recording valid positions at the outset of a new trip segment. This might show up as a gap between the ending of one trip segment and the start of the other (e.g. starting with a distance from home, while it started at home). One solution is to increase the data collection rate; however as discussed in section 2.2.4, this might be a limitation as it highly consumes device battery, device memory, and computation cost. Determining the exact location of indoor activity (which might be significant places to the GPS carrier or origins and destinations) not only is a useful element in interpreting human travel behaviour, but it could also be advantageous for detecting the mode of transport (Stopher, 2008). With multi-day data and a trip that is repeated from day to day, it might be possible to infer the details of a bad-data trip by imputing the details from the same trip made on another day of the week, where a cold start problem did not occur.

2.5.3.2 Identifying the Occurrence of Indoor Activity

Another problem is identifying the occurrence of indoor activity in the first place. As previously described in section 2.3.1.2, entering into tunnels or using the underground train service can be identified by snapping to transport networks; however, indoor activity remains an important problem to resolve. Some solutions might appeal to different types of GPS device categories. In the case of an interactive GPS, respondents could initiate time stamps for when they are at a significant stop. With in-vehicle GPS, time stamps are marked by turning on/off the vehicle engine (Schönfelder & Samaga, 2003). However, not all stops of interest involve turning off the engine. For passive portable GPS, a large proportion of the research community appears to agree on defining a temporal threshold to identify the beginning and end of trips usually known as the “Dwell Time”. This dwell time was agreed to be from 30 to 120 seconds either without GPS signal or of being stationary to define a stop (Schönfelder, et al., 2002; Chung & Shalaby, 2005; Wolf, et al., 2003; Forrest & Pearson, 2005; Bricka & Bhat, 2006; Li & Shalaby, 2008; Bohte & Maat, 2009; Stopher, et al., 2008a).

Some issues with the dwell time concept is that 30 seconds could be too short (e.g. a stop at the traffic lights being identified as a trip end). On the other hand, 120 seconds could be too long (e.g. purchasing fast food/picking up/dropping off a passenger). Brika and Bhat (2006) report that with the shorter dwell time of 45 s, the trip underreporting rate was found to be 31%, whereas it dropped to 12% when the longer dwell time of 120 seconds was used. It might also be useful to bear in mind that most of the previously mentioned studies use an epoch rate of collection of 1 second while for a setting of 60 seconds for example this threshold might prove to be unfeasible.

Stopher et al. (2005) also adds a change in heading to a high dwell time constraint to enable more flexibility yet control on the acceptance criteria for in-vehicle GPS. They still define a 120 seconds dwell time indoors solely to identify a stop/trip end for portable wearable devices.

Ashbrook and Starner (2003) vary the dwell time “t” and plot the number of places found for many values of t on a graph (Figure 2.9). As there is no clear point from the graph to choose, they propose in their study the dwell time to be 10 minutes (600 seconds). The reason to be that it is “safer” than choosing smaller values such as one or five minutes because urban canyons might cause signal loss with re-acquisition times of 45 seconds to five minutes (Garmin Inc., 2011). Furthermore, Schüessler and Axhausen (2009) decide to use 900 seconds as a dwell time for the same reason.

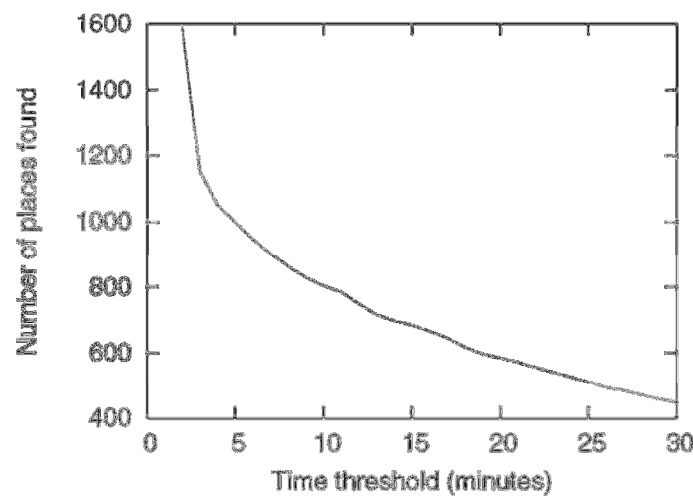


Figure 2.9 Number of places found for varying Dwell Time values for data from (Ashbrook & Starner, 2003)

Liao et al. (2007) use a different approach than the dwell time method. They attempt to take the user’s context information into account. They do that by using an iterative algorithm that re-estimates activities and places after originally detecting them using the data’s temporal information. The temporal information could be such as the time of day, day of week and duration of the stay. For example, time of day can be night, and hence the probability of having a significant stop is high where the user is probably at home, or daytime which would mean at the workplace. This could work for lots of cases; however, not all participants have the same temporal and locational daily patterns.

2.5.4 Discussion on Segmentation Issues

This section discussed the different methods aiming at segmenting a GPS track into trips, stages and stops by respectively identifying mode switches and stops within stages. We have also demonstrated that there are three methods to identify any of these three properties, namely being; clustering, identifying change points and identifying indoor movement.

As for **clustering** GPS fixes, previous studies present a variety of advantageous approaches being simple, with low cost computation, and efficiently identifying the participant’s significant places. On the other hand, these studies bear several limitations such as not

accounting for slow speed trajectories such as walks. Some do not neither account for dense transport networks, which does not make them robust to be used on different networks. Most methods do not possess labelled data to validate the developed algorithms, or in some cases only testing them on a single participant. Another limitation GPS clustering methods possess is not performing spatiotemporal clustering, instead focusing solely on the spatial domain where most of the studies aim at identifying re-occurring stops rather than all stops.

Several studies aiming at inferring the mode of transport from GPS data tend to search for **change points** where the mode changes, breaking the trip into several stages. Some studies base the mode classification stage on this segmentation. The problem with these methods is that the overall classification accuracy is heavily based on the segmentation accuracy. Furthermore, these methods present another type of information-loss where they might classify a low-speed stage incorrectly as it was originally part of a bigger stage with a higher mean speed, or vice-versa. For identifying these points however, some studies use sudden change in speed or acceleration which might produce many errors such as in situations where a vehicle stops at traffic lights or suddenly approaches a traffic jam. Other studies assume that any two stages are always separated by a walk stage, which still might fail in cases of *cycling* or driving the *car* out of a *train* station's car park for example. Others simply restrict which modes can follow one another. An example of that is using a transition matrix between different modes that states probabilities of switching from one mode to another, however only applying this to 4 modes.

We have also highlighted previous studies that attempted to identify static indoor activity. Some studies attempt to identify the exact location of indoor activity, which is a challenge due to the GPS cold start problem. One solution is to increase the data collection rate which has the limitation of highly consuming the device battery, device memory, and computation cost. With multi-day data and a trip that is repeated from day to day, it might be possible to infer the details of a bad-data trip by imputing the details from the same trip made on another day of the week, where a cold start problem did not occur. Another problem that previous research has tackled is identifying the occurrence of the indoor activity. A common practice is to base the search on a threshold "dwell time" period. The value chosen for the dwell time however, will largely depend on the GPS rate of collection. Some suggestions were made to as it is safer to choose a smaller value, such as one or five minutes, because urban canyons might cause signal loss with re-acquisition times of 45 seconds to five minutes. Other approaches also recommend using the time of day and/or other temporal resolutions to identify the most probable destinations the participant would end their trip at.

After identifying clusters, switch points and indoor activity, a long standing argument would be what to count as a stop. Common stops can be counted as any cluster within a road network before or at an intersection or an extended traffic stop, or within any network at bus stops, train stops, etc. However, identifying stops could be confused with longer significant stops since a destination might be located very near to these network features (Stopher, et al., 2005). Therefore, as discussed in subsection 2.3.2.2, a clear distinction and standardisation too of what is to be tagged as stops will lead to accurately assessing the efficiency and accuracy of identifying these stops when validating the attained results.

2.6 Using GIS with Transport Network Datasets

The unlimited amount of geographical data that is available today in the age of huge datasets provides us with endless opportunities of using this kind of data in different applications. With regard to the purpose of this research, the availability of urban transport network datasets bears many advantages in the inference process. In this section we discuss how previous research uses such GIS datasets to relate to GPS data, overcoming GPS error in positioning. We also describe how we use this to the advantage of this research by matching GPS data to their corresponding transport network, leading to detecting the mode of transport. Hence, we introduce the concept of map matching which is well investigated in the literature. We also coin the term “network matching” that was used previously by several research attempts aiming at detecting the mode of transport.

2.6.1 Map Matching

The map mismatch is a problem of lateral movement of position from the GPS trace with respect to the transport network. To overcome this problem, a process called map matching is applied to the GPS data. Map matching is the process of assigning every point to its corresponding network link in a given network. Map matching could hence be defined as the process of snapping the displaced GPS points to the correct road link (Quddus, 2006). Map matching could be advantageous as a GPS data smoothing process that decreases the positional error and hence reduces the distance and speed errors. Therefore, the map matching is also advantageous as a method that helps increasing the accuracy of the modal classification and significant stop identification. A lot of research has attempted to develop algorithms that perform the map matching process, however; with some limitations.

Map matching techniques could be categorized into *vehicle* and *non-vehicle* approaches. This is due to the fact that vehicle trajectories (car, bus, train & tube) have different properties than the non-vehicle trajectories (pedestrians & bikes), in which the former abide by networks and their rules, while the latter are based on different network topology rules influenced by random human behaviour. Most research tackles the vehicle trajectories specifically the car mode, and on the other hand there is little work done on non-vehicle and other multi-modal trajectories (Quddus, 2006; Yang, et al., 2003).

The only studies that tackle other modes attempt to perform the map matching process for multi-modal trajectories with some limitations such as, among others, being tested in non-complex areas or having limited sample size (Ren & Karimi, 2009), or test the process in unusual circumstances such as using very long epoch rates (Yang, et al., 2005). In a multi-modal based trajectories context, the process would not only depend on conventional factors such as the road network, speed, etc.; but will also involve other factors such as geometrical relationships, along with different limitations that will define this process. Other research that attempts to solve a similar problem had very extreme limitations.

Yang et al. (2005) proposes an algorithm that uses topological and geometrical relationships to perform the map matching process, and gives a relatively good accuracy (90%) but does not necessary calculate the right road links but instead it only snaps to the closest nodes. It also only considers tracks with large epoch rates thinned from 1 second data. It also uses a sample

of only 3 routes; therefore, judgement on their choice is not clear in the text which could make a great difference in results.

Ren and Karimi (2009) propose a chain-code-based map matching algorithm which uses a Hidden Markov Model (HMM) for GPS-based wheelchair navigation. It uses the topology of the sidewalks segments of the transport network to carry out the map matching process. It gives an accuracy of (92%) for only 3 tracks that have been tested. Apart from the limited sample size for testing, a collection rate of 1 second was used for testing the algorithm, which is unfeasible in real life due to battery limitations of current GPS devices. The test was also conducted in a non-urban low density area, which does not reflect the typical transport network scenario within an urban area. Among other limitations are that the study is only limited to wheelchair travel mode, completely disregards topology, and the outcome was a “most probable solution” which might sometimes be very uncertain in complex network areas.

Chung and Shalaby (2005) was the only research to our knowledge that proposes a method that attempts merging the map matching and travel mode detection processes. It generates an accuracy of (79%) for links travelled and (92%) for detecting trip modes. However, it only considers one-purpose trips, and it uses 1 second data which is again not feasible in terms of current battery requirements. It also assumes that both the ends of any trip are not in an urban area, and ignores the cold start problem of GPS receivers.

All these attempts attempt to enhance the accuracy of GPS data calculations by snapping to transport networks. However, for multi-modal data, the process is even more complicated since the developed method needs to identify first which network to snap to. Therefore, another type of method needs to be developed to tackle this challenge. The following subsection discusses this approach and previous attempts aiming at tackling this challenge.

2.6.2 Network Matching

The “single-mode” map matching process is not ideal to be used on multimodal GPS data. This is due to the fact that map matching attempts to match GPS to a specified transport network, while the network to be matched to is unknown whether to be the bus, train or underground networks. On the other hand, a method that tests which network GPS data can be matched to is ideal for this research purpose, where the matching is carried out to different networks. In other words, this process would be performing map matching but to different networks. We propose to name this process as the process of network matching, in order to distinguish it from the traditional well-known map matching process.

Apart from selecting the corresponding network to match GPS data to, network matching can also have the advantage of narrowing down the potential modes of transport that a stage can be classified as. As previously mentioned in section 2.5.3 for example, in Australia Stopher et al. (2005) detects whether a participant was using a car or public transport by reasoning using a GIS transport network. The method is based on a comparison of the estimated speed with the actual speed taken to enter and exit the urban canyon. The method does not perform any matching or uses any topological or geometrical relationships between consecutive GPS fixes.

Another study in Toronto uses fuzzy logic- based mode identification algorithms both with and without additional GIS-based analysis (Tsui & Shalaby, 2006). The study achieves a rate for accurate mode detection of 91% illustrating that using network data improved the bus detection rate from 76% to 80% by taking advantage of the bus route information. In addition, matching GPS traces with public transportation routes in GIS made it possible to identify 88% of the streetcar travel and 100% of the subway travel, yielding a detection rate of 98% for off-road modes.

Bohte and Maat (2009) achieved a 70% success rate in a mode detection study carried out in the Netherlands, with 75% for car mode, 35% for rail, 72% for bicycle, and 68% for walk. One limitation was that the bus mode was not included. Moreover, the assignment of a mode was almost exclusively based on average and maximum speeds of trips, although network data was used to separate rail from car mode.

To the knowledge of the author, Gong et al., (2012) is the only attempt to solely use network matching to detect the mode of mode of transport from GPS data. The study reports 82.6% success rate after testing the algorithm on GPS trips in New York considering all modes of transport. However, the study does not use routing information, uses no sampling technique, ignores seasonal variation by using 1-day data, is based on a pre-segmentation process, and uses a very basic clustering technique which was not assessed individually.

2.6.3 Summary of Methods using GIS with Transport Network Datasets

This section discusses the attempts to use GIS and transport network datasets to enhance the accuracy of GPS data and to infer the mode of transport. We describe the concept of map matching which is the process of assigning every point to its corresponding network link in a given network. Methods that use map matching apply it for smoothing GPS data decreasing the positional error and hence reducing the distance and speed errors. Therefore, the map matching is also advantageous as a method that helps increasing the accuracy of the modal classification and significant stop identification. Most research in this area only tackles the vehicle trajectories specifically the car mode. The few methods that apply map matching to multi-modal trajectories use a variety of successful techniques to achieve good accuracies. However, the validation techniques used are among the main limitations of such methods where they use limited sample sizes and sample compositions that are not based on any statistical evidence. Other limitations include using a very short GPS data collection rate which is unfeasible due to battery and memory limitations of GPS devices and developed algorithms.

In this section, we also coin the term network matching for methods that aim to select the transport network that a GPS fix belongs to. Not only these methods have the advantage of selecting which network to snap to in the map matching process, but some of them also have the advantage of detecting the mode of transport from the GPS data based on this selection. On the other hand, most of these methods limit the network matching process for differentiating only one mode from the rest. Others do not use any topological or geometrical relationships between consecutive GPS fixes, which are highly advantageous in understanding the overall movement of the trajectory on the transport network, and hence, enhancing the accuracy of the mode detection. None of the studies performs the network matching using

routing information, which means that the matching does not test whether trajectories follow unique routes such as bus or underground routes. This in turn means that the matching is only based on the distance to the network, whereas a trajectory might be jumping from one line to the other. There are also many sampling issues similar to the limitations described in section 2.2.

2.7 Summary and Conclusions

This chapter discussed the different issues affecting the accuracy of the studies attempting to detect the mode of transport from GPS data. As previously mentioned, these issues can be categorised into **validation** and **method** issues. Figure 2.10 illustrates this categorisation, where validation issues affect the process of calculating the accuracy while the method issues relate to the efficiency of the adopted technique to carry the classification out. The validation issues relate to the primary phases of dealing with GPS data, namely being; the *data collection* and *data pre-processing* phases. On other hand, the method issues are divided into different types of attempts to detect the mode of transport. One type is work that uses *classification* technique in order to distinguish between different modes such as machine learning techniques. Another type of studies attempts to *segment* the GPS track into different segments (stops, stages, and ends). A third type attempts to use *GIS* to match GPS fixes to transport networks, and hence enhancing the GPS data positional accuracy and detecting the corresponding mode of transport. These three categories of studies sometimes are combined together in different ways to construct a framework that benefits from each.

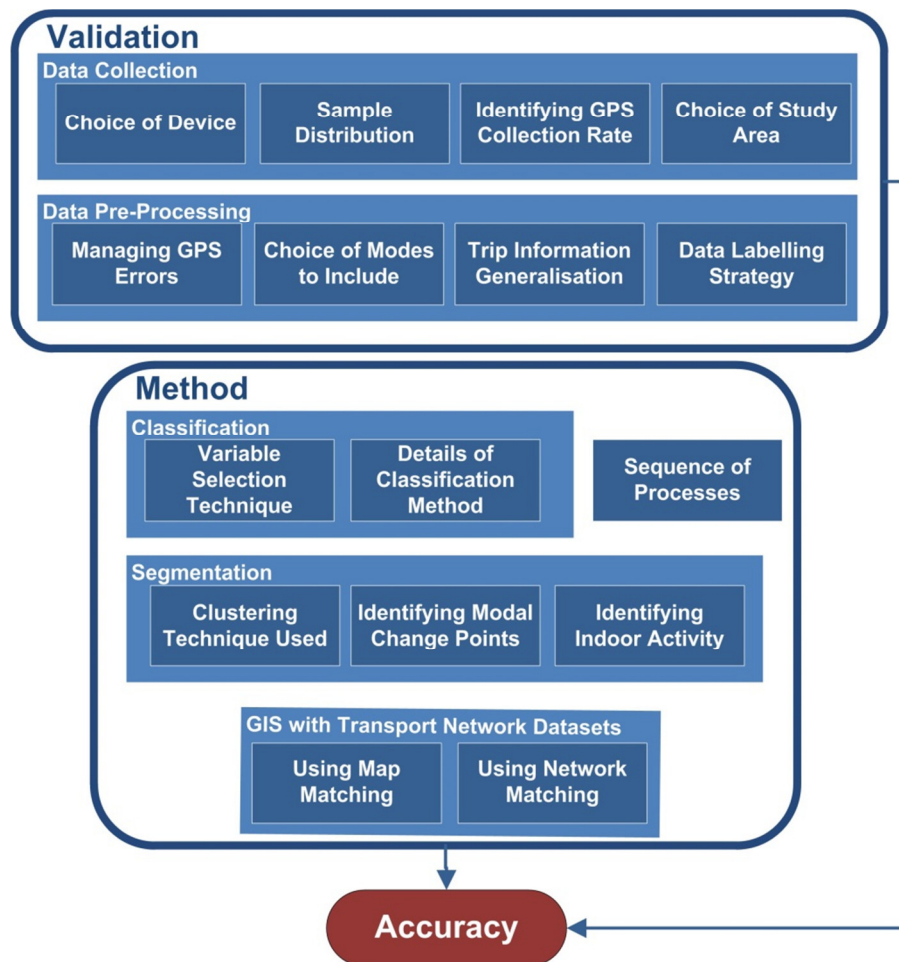


Figure 2.10 Issues affecting Accuracy of Methods detecting Transport Mode from GPS Data

2.7.1 Validation (Data-Related) Processes

We first described the method validation issues that stem from how GPS data is collected and pre-processed. Regarding the **data collection** process, we have discussed several issues such as the device type used, study area, sample distribution size and duration, and sample temporal granularity. We have demonstrated that the *device type* used in previous studies depends on what the study aims to collect and on the environment of study area. Among these properties is having a long battery life, small size, good accuracy, providing certain coverage, being interactive, or having real-time capabilities. As for the *study areas* where previous studies have been testing their algorithms, there has been often limitations such as being very simple or having data not covering many areas of its transport networks. *Sampling* issues also include that most studies did not provide evidence that their samples had a representative coverage of the study area. They neither account for the weekly seasonal variation which creates a bias towards that specific week of collection. Moreover, none of the studies which calculate the sample size and duration for GPS-based travel surveys take into consideration the modal, spatial or temporal granularities of the designed study. Another sampling problem is defining the **GPS data collection rate**, where a very fine rate of collection holds limitations. These limitations include overloading the memory and battery power of current GPS devices or smart phones, as well as adding to the computation cost of any of the used algorithms which gives rise to problems like increasing the burden on the participants to charge their devices very frequently and act as a constant reminder that they have a tracking device. Not much research provides detailed analysis of the best GPS epoch rate setting for tracking participants within an urban environment. Moreover, all studies aiming at inferring the mode of transport from this data tend to use the finest grained temporal granularity for collecting their data without any statistical or logical justification.

Regarding the **data pre-processing** process, we have described two types of pre-processing stages; one of which is dealing with GPS errors and the other being adding metadata to the collected GPS data. For *dealing with GPS errors*, we have explained the guidelines and practices previous research used to dealing with and benefiting from GPS limitations by filtering and processing GPS data. We demonstrated that there are two cases of implications of GPS technology limitations, namely; low positional accuracy and signal loss. *Low positional* accuracy caused problems such as wandering errors due to environmental factors such as urban canyons. These were treated through filtering data by dismissing fixes with less than 4 satellites, HDOP more than a value of 5, unrealistic altitude value, speeds above 50 m/s or accelerations above 10 m/s² according to different congestion zones or transport networks. Other studies apply smoothing techniques for eliminating urban canyon effects or spatial clustering for wandering effects. Spatial clustering is also used in segmentation by identifying stops. As for *signal loss*, it is categorised into partial and total signal blockage. Partial blockage was treated by integrating the GPS device with other devices such as gyros, Wi-Fi, or GPRS technologies. The disadvantage of these hybrid solutions is that such equipment is usually expensive, less portable, or consumes high battery power. On the other hand, total blockage has been treated by using transport networks to find missing segments of a GPS track. Total blockage has also been used to detect indoor activity. Total blockage could be also beneficial for identifying the certain modes such as underground travel.

As for the addition of *metadata to the GPS data*, some decisions made regarding this information hold many limitations. Among these limitations is considering a *limited number of modes of transport* for the developed method. This makes these methods less robust to identify ignored modes. Another limitation is that some studies *generalise the modes* used in a trip to only the most dominant mode within this trip. This decreases the accuracy of the learning and validation process as a result of having a mode being denoted by mixed modes. The metadata, such as the mode of transport, used to validate the results of the inference method is usually added by the participants reporting the details of their GPS tracks. Some studies do not use participant-reporting to add this metadata to the GPS data track and instead depend on assumptions that may bear a large amount of inaccuracy. These studies do not collect participant feedback due to reasons such as the being burdensome and bearing privacy issues for the participant bears, as well as being slow and expensive for the data collector. Instead, some web-applications exist that enable the participants to view their tracks and edit them, yet still bearing several limitations such as not being able to add data metadata. Only one study enables the participants to fully check and edit their tracks. However, the study holds other limitations such as presenting unfamiliar basemaps, not using useful web-mapping products useful for better recall such as Google Streetview, and enabling the user to only use the data the data collector provides limiting the interaction of the participants with the data.

2.7.2 Classification Method-Related Processes

As shown in Figure 2.10, issues related to the type of method used for the detection of the mode of transport from GPS data are the second contributor to achieving a high accuracy. As we described in this chapter, there are three types of methods by which studies attempt to detect the mode of transport. The issues/shortcomings within these three areas are as follows:

First, we discussed in this chapter methods based on **classification** techniques, where these methods depend on a technique to classify each GPS segment into a mode. These techniques varied from machine learning to procedural approach. A major limitation is the lack of any statistical evidence for the choice of the IV (or group of IVs) to be used for the classification in order to achieve the highest accuracy. This limitation results in the usage of IVs that are not the most discriminate between the different classes (modes of transport), and hence, not achieving the highest possible accuracy. The impact of using different techniques for classification on the achieved accuracy is also very high due to the difference in the nature of the used methods.

Problems with procedural approaches include the lack of considering the sequence of GPS segment strings together as one continuous chain of movements. On the other hand, limitations of ML methods include the absence of the ability to learn and adapt after solving a problem as some expert systems can, being extremely sensitive to small perturbations in the data which results in over-fitting or having a problem carrying out an out-of-sample prediction. Despite SVM's limitation of being computationally slow, yet it appears to benefit solving the classification problem under discussion in this research for various reasons. Among these advantages is that SVMs can be easily trained and require no feature extraction, leading to learning efficiently from the data structure, which is useful for understanding trajectory movements in different modes. Another common problem is the usage of temporal

information for classification which does not account for participants with abnormal schedules.

Second, we discussed **segmentation** methods that break a GPS track into stages and stops by respectively identifying mode switches and stops between and within stages. We have also demonstrated that segmentation has been conducted in travel survey methods using either clustering, identifying change points and identifying indoor movement. There have been many limitations with these three types of methods. Among *clustering* limitations is not accounting for slow speed trajectories such as walks or for dense transport networks, leading to a lack of being robust to be used on different networks. Most methods neither possess labelled data to validate the developed algorithms, or in some cases only testing them on a single participant. Another limitation GPS clustering methods possess is not performing GPS sequential spatiotemporal clustering; instead, aiming at identifying significant stops to the participant rather than stops.

As for methods aiming at *identifying modal switch points*, the main limitation is that the overall classification accuracy is heavily based on the segmentation accuracy where a small stage might not be as representable of a larger one, and vice-versa. Some studies use sudden change in speed which is not useful for situations such as when a vehicle stops at traffic lights or suddenly approaches a traffic jam. Other limitations include assuming that any two stages are always separated by a walk stage, which is not always true in certain situations. Therefore, some studies restrict which modes can follow one another. An example of that is using a transition matrix between different modes that states probabilities of switching from one mode to another, however only applying this to 4 modes.

For *identifying static indoor activity*, some studies increase the data collection rate to eliminate the cold start problem, which has the limitation of highly consuming the device battery, device memory, and computation cost. However, other studies tackled the cold start problem by studying repetitive daily trips. Another problem that previous research has tackled is identifying the occurrence of the indoor activity. Many studies base the search on a threshold “dwell time” period, however, this largely depends on the GPS rate of collection. Some suggestions were made to as it is safer to choose a smaller value, such as one or five minutes, because urban canyons might cause signal loss with re-acquisition times of 45 seconds to five minutes. Other approaches also recommend using the time of day and/or other temporal resolutions to identify the most probable destinations the participant would end their trip at.

After identifying clusters, switch points and indoor activity, an important issue is what is to count as a stop. And the answer to that would depend on the amount of trip detail required. Hence, a research that aims at detecting these segmentation features needs to first clearly define what is to be considered as a stop. That will lead to a clearer and more accurate segmentation, and will minimise problems such as confusing spending 5 minutes indoors in an underground station with reaching an indoor destination.

Third, we discussed studies that carry out **GIS** attempts to contribute to the problem of detecting the mode of transport using **transport network datasets**. These methods *use GIS* to enhance the data accuracy, identify a certain mode, or directly classify GPS data into modes.

Initially, some methods use *map matching* to enhance the accuracy of GPS data by snapping to the closest transport network. These methods apply map matching to smooth GPS data decreasing the positional error and hence reducing the distance and speed errors. Many of these methods only tackle the vehicle mode, which is a major limitation in these studies. Most of the remaining limitations include are in the way that these methods validate their results. Other limitations include using a very short GPS data collection rate which is unfeasible due to battery and memory limitations of GPS devices and developed algorithms.

Other methods use *network matching* to identify the transport network that a GPS fix belongs to, and hence, detecting the mode of transport. However, most of these methods limit the network matching process to one mode. Others do not use any topological or geometrical relationships between consecutive GPS fixes, which are highly advantageous in understanding the overall movement of the trajectory on the transport network, and hence, enhancing the accuracy of the mode detection. Other limitations include the lack of using routing information, leading to the allowance of a big source of matching error. This in turn means that the matching is only based on the distance to the network, whereas a trajectory might be jumping from one line to the other.

Chapter 3

Framework of Methodology

3 FRAMEWORK OF THE METHODOLOGY³

In chapter 2, we have demonstrated the importance of understanding travel behaviour using travel surveys as a means to collect travel information. We have also highlighted the importance of inferring trip information from positional data resulting from GPS-based travel surveys. This information extraction from GPS trails eliminates many of the problems that travel surveys suffer from. The inference of trip information from the data minimises the financial and time cost that these surveys incur, as well as minimising the participant burden triggering an increase in response rates (Bricka & Bhat, 2006).

One important information type that many studies attempt to infer from GPS data is the mode of transport. This inference could largely replace or complete a lot of the feedback required by users when labelling and tagging travel diaries. As highlighted in chapter 2, although these studies have proposed solutions to different parts of problems that face this type of inference, yet many of these studies still suffered from problems with the validation model or the method adopted.

Figure 2.10 in chapter 2 summarised the range of issues in previous attempts to infer the mode of transport from GPS data in different stages of the process. As highlighted in the previous chapter, the validation limitations are mostly problems related to the quality and characteristics of the data. On the other hand, the method limitations relate to limitations in the framework of developed methods that were used in the past. In this chapter, we discuss the framework we develop to infer the mode of transport from sparse GPS data, while attempting to address all the validation and method limitations that we have mentioned in chapter 2. Since these limitations were separated into two separate parts in the literature review, the framework is similarly broken into two parts. The first part defines, develops and standardises methods by which we assure maintaining high data quality. The second part, being the main body of the research methodology, develops a sound method to infer the mode of transport using a combination of methods that tackle different issues that challenge understanding the movement of a trajectory.

This chapter is divided into several sections. The first provides a full description of the framework with its 5 phases discussing how each part of each phase addresses certain limitations highlighted in the previous chapter. The second section describes how this framework addresses the aim and objectives specified at the beginning of this thesis in chapter 1. The third and final section of this chapter discusses the structure of the rest of the thesis based on the 5 phases of the developed framework along with the validation and discussion of results.

3.1 Framework Description

In chapter 2, we highlighted the limitations that previous methods aiming at detecting the mode of transport from GPS data possessed and opportunities that some techniques offered that could be harnessed to feed into the detection process. In this thesis, we use these limitations and opportunities to investigate and construct a complete process that detects the

³ Some parts of this chapter is based on a previous publication of ours: Bolbol, A, Cheng, T, Tsapakis, I & Haworth, J, 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems*. (In Press)

mode of transport from GPS data starting from the data collection phase to the method used to resolve this classification problem. We have also highlighted in chapter 2 that the issues involved in this process can be broken into two general categories, namely; validation (data-related) issues and method-related issues. This is illustrated in Figure 3.1 reflecting both categories that this thesis will present in the chapters to come; broken down into different phases that constitute this thesis. From the figure, the first category consists of two phases, namely; data collection and data pre-processing. This category generally attempts to standardise the data-related issues in a way that maximises the data quality level and validation process. The second category presents our attempt to detect the mode of transport from sparse GPS data by developing a method that consists of three phases, namely; classification, segmentation and network matching. This section describes these two categories, specifically the latter, emphasizing the structure of the developed method in this research.

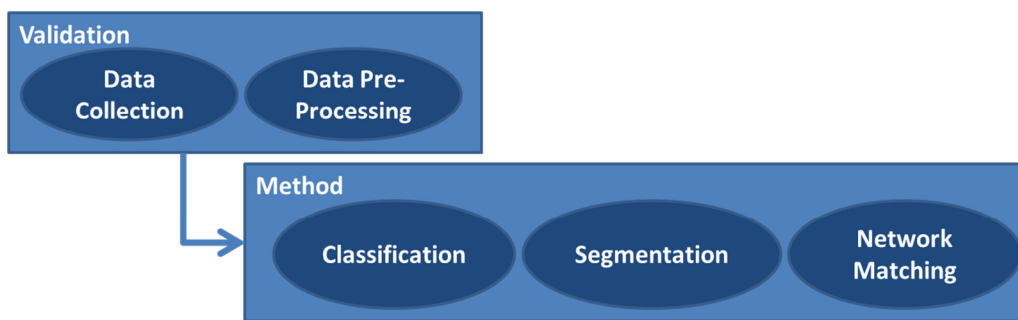


Figure 3.1 Phases of Thesis based on Limitations & Opportunities mentioned in Chapter 2

3.1.1 Validation (Data-Related) Processes

In chapter 2, we first discussed the data-related limitations which include the data collection and the data pre-processing issues (as shown in Figure 3.1). In this subsection, we describe our efforts in this thesis to tackle limitations or lack of guidelines in data-related issues in the context of GPS collection for mode detection studies.

3.1.1.1 Data Collection

We first start by tackling the first phase (data collection) in order to obtain high quality data using the most optimal data collection specifications that help minimise the effort and cost involved in this process. The *data collection-related* issues and limitations included the selected sample composition of the study, granularity of the data, choice of data collection device, and the choice of the study area.

We have first discussed the sample composition issues, shedding some light on the issue of the sample composition previous studies have used for testing their classification algorithms developed to infer the mode of transport from GPS data. However, not many of these studies looked into assessing the spatiotemporal distribution of the sample across the study area in order to provide an appropriate validating of these techniques. Therefore, in this thesis, we describe the sampling spatio-temporal distribution and characteristics as a way to inform the adequacy of the sampling strategy for different modes of transport of different spatial analysis

zones for different temporal granularities. Trips' origins and destinations are also analysed in order to identify of activity spaces across central, inner and outer London.

Another highlighted sampling problem discussed in chapter 2 is defining GPS collection rates (temporal granularity) by which previous studies have collected their data. While more data is usually encouraged, yet it is unnecessary for this kind of research as it adds to the battery and memory consumption of existing GPS generations available nowadays in the marketplace, while adding to the computational cost of processing vast amounts of redundant data. Therefore, we also investigate the optimal collection rate required for such travel diary GPS data collection.

As mentioned in chapter 2, another issue in the data collection process is the device used for collecting positional data. We have mentioned that existing devices possessed different properties such as having a long battery life, small size, good accuracy, providing certain coverage, being interactive, or having real-time capabilities. In this thesis, we describe the effect of these properties on the positional data collection process, and we attempt to select the device with the best configuration to collect positional data which is best fit for detecting the mode of transport.

The study areas chosen for testing such algorithms also often have limitations such as being very simple or having bad data coverage in many areas of its transportation networks. Therefore in this thesis, we also discuss the basis of the selection process of the study area to conduct the intended study.

3.1.1.2 Data Pre-Processing

Chapter 2 also discussed problems that arise when processing GPS data, and demonstrated how some research handles it. One specific problem is GPS data coverage limitations that are exhibited in two forms; namely, low positional accuracy and signal loss. In this research, we deal with low positional accuracy by first filtering data according to specific thresholds of parameter values. We also develop an innovative spatial clustering procedure that identifies committed stops within the track of a GPS trajectory, leading to the elimination of the wandering effect that GPS inaccuracies result in. On the other hand, total blockage can be as a result of underground travel or going indoors. In this research, we use total blockage in our methodology of inferring modes of transport using a couple of procedures. First, for underground travel, we use transport networks to identify whether a trajectory matches to one and obeys its topological rules, to make an inference of the mode used (e.g. train). Second, for indoor instances, we develop a segmentation procedure as part of our methodology that reasons on whether a trajectory is indoors or simply suffered a temporary signal loss.

Another type of issues is that related to the quality of validation. As discussed in the previous chapter, the validation process is as accurate as the quality of the metadata associated with the data. This metadata quality largely depends on the decisions made at the data pre-processing stage relating to topics such as the types of mode of transport that are to be investigated, the generalisation level of the trip information, and the strategy used to report the mode of transport by the participants.

In this thesis, we decide on the most common maximum modes of transport to include in this study from previous work, taking into account the nature of many modern cities that exist around the world. On the other hand, the level of detail a GPS track is to be broken into is of the most influence on the validation process. Among the limitations of previous literature was setting a too low level of generalisation of a track and using a loosely defined level of track elements. In order to overcome these limitations, we define elements that constitute a GPS track to a highly detailed level, so that labels, origins and destinations of modes of transport can be assessed and identified efficiently. We also define the segmentation rules by which a track is to be broken down into according to different use case scenarios. Scenarios are also developed to identify stops within the track, whereas the situational accuracy of defining these stops feeds largely into the accuracy and standardisation of the validation rules by which a classification method is assessed. Standardising and properly defining the stops identification process also leads to a more accurate classification of elements of a track.

Another issue identified in research is the metadata pre-processing influence on the validation process. The validation process of the developed classification algorithms are only as good as the metadata attached to the GPS track is. As demonstrated in chapter 2, many previous studies attempting to infer the mode of transport from GPS tracks do not use participant-reporting to add metadata such as the mode of transport to their tracks. Instead, these researches depend on assumptions that may bear a large amount of inaccuracy to minimise cost and time. Other research that attempts to collect this metadata from the GPS participants usually faces problems such as low response rates or inaccuracies due to factors mentioned in chapter 2. In this thesis, we develop a web interface that enables users to edit their tracks and add metadata such as the mode of transport to overcome these problems. The interface takes advantage of the Geoweb 2.0 technologies and crowdsourcing user attitudes that exponentially increase volunteered web activity to attract participants to contribute edits and metadata to their tracks in a timely, accurate, user-friendly, and interesting fashion. The interface also uses a familiar basemap (Google Maps) with Google Streetview as an extended option, thus facilitating track details recall.

3.1.2 Classification Method-Related Phases

The main focus of this research is to infer the mode of transport from sparse GPS data. Therefore, the second part of this thesis discusses our attempt to achieve this inference using a framework that proves efficient, achieves high accuracy and overcomes method limitations from previous research studies. Figure 3.2 summarises the framework developed in this research to infer the mode of transport from sparse GPS data.

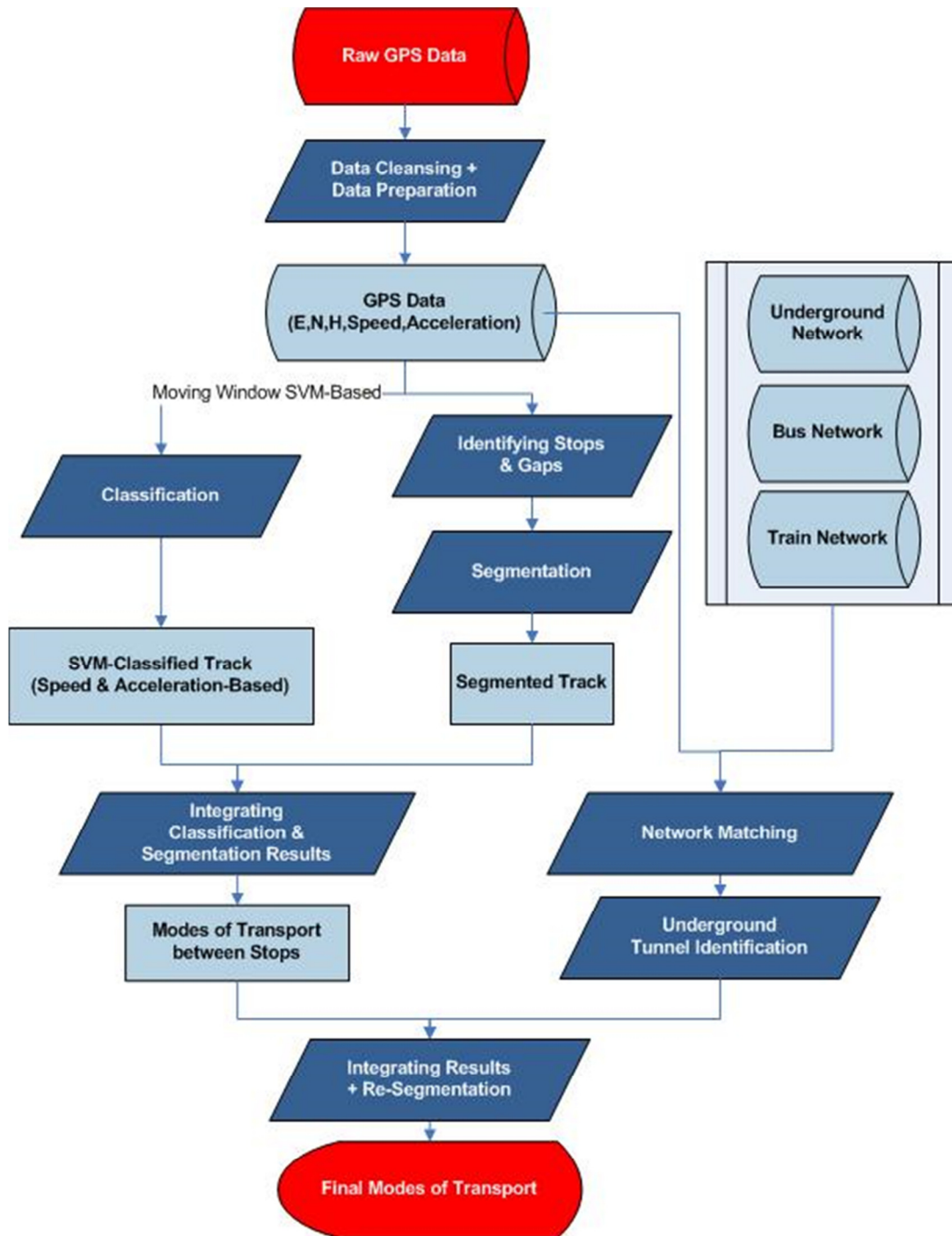


Figure 3.2 Framework produced in this Research to infer Transport Mode from Sparse GPS Data

The inference framework consists of several phases that are summarised in the right part of Figure 3.1. These phases are start with a classification and segmentation phase in parallel, and are followed then by a network matching phase that verifies the achieved results using GIS datasets. As demonstrated in Figure 3.2, the framework starts by cleansing the raw GPS data as demonstrated in the beginning of section 3.1.1.2 by filtering the data according to specific thresholds of parameter values. The data is then prepared as coordinates, speed and acceleration values to qualify to enter into the framework. The dataset then enters the first two phases in parallel, namely; the classification and segmentation phases. The classification

phase uses a moving window Support Vector Machines based algorithm using the speed and acceleration values to output initial mode of transport labels for each segment of the GPS track. On the other hand, the segmentation phase uses an innovative spatial clustering algorithm that identifies stops and gaps in the GPS track. The output from both phases then are integrated into one single output containing a track that contains initial mode of transport labels and that is also segmenting the track into GPS stages, each stage consisting of a single mode of transport. This single output then enters into a two-step process of network matching. The first step of the network matching process identifies underground travel using the London Underground tube network using a reasoning algorithm using the underground network dataset provided. The second step uses the underground, bus and train network datasets to identify verify results from the initial classification. The results are finally integrated together and a second step of logical reasoning is applied to produce the adjusted modes of transport as a final output. The rest of this subsection discusses these three phases in further detail.

3.1.2.1 Classification

As demonstrated in chapter 2, one major limitation of all classification methods is the lack of strategy or statistical evaluation for the choice of the Independent Variables (IV) to be used to differentiate between different modes of transport. Another issue is the impact of the classification technique used for classification where most procedural methods for example contain problems such as ignoring the sequence of GPS segment strings together as one continuous chain of movements, resulting in losing movement patterns of a trajectory. On the other hand, Machine Learning (ML) methods have different problems depending on the technique used. *Fuzzy logic* for example does not have the ability to learn and adapt after solving a problem as some expert systems can. *Decision trees* can be extremely sensitive to small data perturbations leading to over-fitting, as well as not being as smooth as other ML methods.

To address the classification problems, part of this thesis attempts to fully understand and account for these limitations in the process of inference. We aim to solve a classification problem of GPS data into different transportation modes (*car, walk, cycle, underground, train* and *bus*). This phase is based on Support Vector Machines (SVMs) to classify GPS segments into respective transportation modes. An advantage of using SVMs over other ML methods is that they can be easily trained and are applied directly to the data without the need for a feature extraction process. This allows us to learn from the structure of the data. The proposed method uses a moving window across every group of consecutive segments in order to capture the nature of participants' movements through different transportation modes. We consider all the possible transportation modes, while testing the algorithm to avoid any mode aggregations or exclusions. An ANOVA analysis is also conducted to identify the Independent Variables (IV) that have the most significant effect on discriminating between different modes of transport. These variables are identified to be speed and acceleration and are to be used in the classification. A segmentation process is applied to the classified data after the initial SVM inference is performed to avoid the reliance on the segmentation accuracy if we have had applied the segmentation before the classification. We also avoid using any temporal assumptions to ensure the robustness of our algorithm over different samples. A transition

matrix is also applied to assign modes in the case of potential transitions between any two non-walk stages.

3.1.2.2 Segmentation

A line of research we discussed in chapter 2 was aimed at segmenting a GPS track into trips, stages and stops by respectively identifying mode switches and stops within stages. In chapter 2, we have discussed three methods to identify any of these three properties, namely being; clustering, identifying change points and identifying indoor movement.

Identifying clusters

Among the advantages of **clustering** techniques in previous studies are that some are simple, require low cost computation, and efficiently identifying the participant's significant places. However, among the limitations of many is the lack of accounting for slow speed trajectories (walk) and/or dense transport networks. Another clear limitation of these studies is developing spatial clustering techniques to find significant places rather than developing spatiotemporal clustering techniques for movement data. Moreover, a lack of labelled data availability to validate the developed algorithms reduces the performance validity of many of these algorithms.

In this thesis, we introduce our approach to detect stops within a GPS track, which clusters GPS points of a moving trajectory based on travel speed and distance. The algorithm reduces the time cost of finding these clusters since it performs basic search in the spatio-temporal sequence domain rather than searching for all nearest points only in the spatial domain as performed in Ashbrook and Starner (2003). The algorithm also identifies a centroid for each cluster presenting the most probable location that the GPS trajectory was present at. We also use the dwell time concept to identify indoor stops in the case of loss of signal. This in turn helps segmenting a GPS track into individual trips. Our approach also overcomes the shortcoming of previous research of not having tagged data to validate results by testing the algorithm on the labelled data collected in this research.

Identifying change points

Generally, studies aiming at inferring the mode of transport from GPS data tend to search for **change points** where the mode changes, breaking the trip into several stages, having many basing the whole algorithm on this segmentation. Problems with these methods include the dependency of the classification accuracy on the segmentation accuracy, which might not be highly accurate in some instances. Another problem is the over-segmentation of some algorithms that consequently leads to over-classification. Moreover, using sudden change in speed or acceleration to identify these points, as some studies do, produces many errors as demonstrated in chapter 2. Assumptions such as that any two stages are always separated by a walk stage, might also fail in cases of *cycling* or driving the *car* out of a *train* station's car park for example. On the other hand, some methods are advantageous such as using a transition matrix between different modes, however only applications of this have been only introduced to no more than 4 modes.

In this thesis, our classification algorithm, described in subsection 3.1.2.1, does not base the inference on segmentation, avoiding many of the highlighted limitations. Rather, the algorithm independently classifies the track and carries out a segmentation process in parallel, then integrates both results as a result. Adopting this approach, the classification results are used to enhance the segmentation results and vice-versa. Moreover, a transition matrix is also adopted after classification to reason about the sequence of modes within a trip.

Identifying static indoor activity

Chapter 2 also shed some light on attempts to identify **static indoor activity**. Solutions such as increasing the data collection rate possess many limitations including high computation cost. Other solutions, such as using multi-day data to infer these locations from repeated trips over several days, seem to achieve good results. However, to judge whether an indoor activity has occurred in the first place is the bigger challenge. The usage of the dwell time concept is a common practice for identifying these occurrences, however, a major problem that arises is the differentiation between indoor activity and travel with no coverage (such as in underground tunnels).

Therefore in this thesis, we use transport networks to differentiate between these instances. The following section describes our attempt to carry this reasoning out. While for identifying indoor activity, we adopt the dwell time concept coupled with a reasoning framework based on temporal and classification attributes.

3.1.2.3 Network Matching

A final group of issues discussed in chapter 2 relate to the usage of GIS and transport network datasets to enhance the accuracy of GPS data and to infer the mode of transport. We have introduced the concept of map matching and its usage in previous research to increase the positional accuracy of the GPS fixes, and hence the accuracy of the modal classification. Very few methods apply map matching to multi-modal trajectories to enhance inference accuracy. However, the limitations of these methods vary from lacking sufficient data samples to validate the developed methods to using a very short GPS data collection rate which is very computationally costly. We have also coined the term Network Matching (NM) in chapter 2 for methods that aim to select the transport network that a GPS fix belongs to. Among the advantages of NM is selecting which network to snap to in the map matching process, and even sometimes detecting the mode of transport from the GPS data based on this selection. However, among the limitations of methods that use NM is that they limit it to only one mode. Another limitation is not using any topological or geometrical relationships between consecutive GPS fixes, which are highly advantageous in understanding the overall movement of the trajectory on the transport network, and hence, enhancing the accuracy of the mode detection. No routing information is ever used neither, which means that the matching does not test whether trajectories follow unique routes such as bus or underground routes. This in turn means that the matching is only based on the distance to the network, whereas a trajectory might be jumping from one line to the other.

In this thesis, we present a novel Network Matching (NM) approach for testing whether GPS trajectories are travelling on specific transport networks. These transport networks include the underground, train and bus networks. We develop NM algorithms that will act as a verification phase of the initial transport mode classification from the previously described SVM classification method. This stage is the final major phase of the whole classification framework verifying if certain classified transport modes actually follow their corresponding network. For example, if a stage (a group of GPS segments) was classified as train, the NM phase will check whether all the GPS segments within that stage are within a pre-specified distance from the train network and follow a specific train route. The network matching process can also be used for several purposes such as a pre-map matching step (To provide prior knowledge of which network to snap to) (Quddus, 2006), analysing different network usages (or network split) (TfL, 2009b), or detecting the mode of transport which is the very purpose of this research. We identify underground (tunnel) travel, however, using an innovative approach that uses a defined dwell time value while matching the travel to underground network stations. The results of this verification phase are integrated with the output from the first two phases producing a final output that is a GPS track that is divided into stages, each containing its respective mode of transport.

3.2 Addressing Research Aim and Objectives

The framework developed in this thesis aims to address the aim and list of objectives set to achieve this aim. This thesis addresses the aim of this research, mentioned in chapter 1, by developing a framework that leads to the segmentation of a GPS trip into stages each classified into a certain mode of transport. This research also standardises and regulates the methods of GPS data collection and data pre-processing when using travel diaries to provide a reliable validation process of the developed framework and a supreme data quality.

To fully address this aim, the four objectives stated for this research in chapter 1 are addressed within the framework developed in this thesis. First, standardising and defining the best practice for location data collection and processing are addressed within the first two phases of this thesis grouped under the validation-related part of this thesis described in subsections 3.1.1.1 and 3.1.1.2. This segment of the work produced, within the context of research attempting to identify the mode of transport from GPS data, assures obtaining high quality data quality and a reliable framework performance validation process.

The second objective mentioned in chapter 1 is concerned with breaking the GPS track into stages, each of a certain mode of transport. This is addressed by combining the outputs from the classification (subsection 3.1.2.1) and segmentation (subsection 3.1.2.2) phases using a reasoning framework to group segments of one stage together.

The third objective concerned with detecting the mode of transport within each stage is addressed mainly using the classification phase followed by the rest of the framework. The segmentation and network matching (subsection 3.1.2.3) phases are considered as complimentary phases to enhance the performance and verify the results of the classification algorithm.

Finally, the fourth objective mentions using the transport network datasets to validate the results from the classification and segmentation phases, and enhance the achieved classification accuracy. We address this objective by developing a network matching algorithm that matches GPS segments to transport networks in an attempt to validate the achieved classification results. The next section of this chapter describes the structure of the rest of this thesis in the context of addressing each of these research objectives individually. The next section also describes how the thesis structure shall present the methods developed to integrate results from each of these phases with one another.

3.3 Detailed Thesis Research Methodological Structure

As previously mentioned; the issues to be addressed by this thesis consist of five phases and can be grouped into two categories as shown in Figure 3.1. The first category addresses the validation-related issues summoned in a data collection and a data pre-processing phase. The second category addresses the method-related issues by developing a mode of transport detection framework consisting of classification, segmentation, and network matching phases. In this section, we describe how each of these phases is handled by an individual chapter describing our attempt to address the aim and objectives of this research.

Chapter 4 describes our attempt to address data collection issues in the context of GPS-based travel survey studies attempting to infer the mode of transport from the collected data. Therefore, chapter 4 starts by describing positional data collection specifications for the purpose of detecting the mode of transport. It describes then properties of existing devices used for positional data collection, and attempts to select the device with the best configuration to collect positional data which is best fit for detecting the mode of transport. The work then explores the basis of the study area selection process to conduct the intended study. Chapter 4 then presents a pivotal study that would define and standardise the sampling technique of similar studies. Hence, the chapter describes our attempt to calculate the minimum adequate sample participant size and study duration to collect GPS data for the purposes of this study. This study ensures the testing collected sample adequacy by understanding its spatiotemporal distribution across the study area during different times of the day. Finally, chapter 4 describes a study we conduct that attempts to calculate the most appropriate rate of GPS data collection which minimises the battery consumption and the computational cost required for processing and analysing data from such studies.

Chapter 5 describes our attempt to address data pre-processing issues raised in chapter 2. The chapter first addresses the GPS data accuracy limitations such as low positional accuracy, signal loss problems and the modes of transport selected to be studied within the scope of this research. This data filtering process shall lead to obtaining the optimal data quality to be accepted for processing in this research. Chapter 5 then defines the level of detail that exists in GPS tracks' information (or metadata), focusing specifically on the level of information granularity we will be assessing in this research. The work also attempts to standardise the process of attaching metadata by the participants when labelling their tracks. This working process standardises the metadata (mode of transport) which the classification method will be tested and validated based on. Finally, the last part of chapter 5 discusses reasons why traditional travel diaries return low response rates or inaccurate information from the side of participants taking part in these surveys. The quality and quantity of the feedback information provided by the participants affects the quality of the method and validation process of the classification framework developed in this research (and any other research). Therefore, we introduce an online web-based interface through which participants can upload, view, and edit their tracks overcoming all the travel diary limitations that provoke low and inaccurate response rates.

Chapters 6, 7, and 8 present our mode detection framework detailed into the phases previously mentioned at the start of this chapter, namely; classification, segmentation, and network matching respectively. In **chapter 6**, we aim to solve a classification problem of

sparse GPS data into different modes of transport (*car, walk, cycle, underground, train and bus*). Our approach attempts to obtain an in-depth understanding of the aspects that influence the classification process. Therefore, we first study the variables that could contribute positively to this classification, and statistically quantify their discriminatory power using an ANOVA analysis. We then introduce a novel approach to carry out this inference using an algorithm based on Support Vector Machines (SVM) classification, mentioned earlier in this chapter. SVMs enable us to account for sequence of movements of a participant rather than each movement individually, and hence, achieving a better classification. We achieve this by using a moving window that classifies instances of data consequent blocks. We then complement this by using logical filters that apply a transition matrix between different phases of the trip. The results of the classification are presented at the end of the chapter, comparing it to results from applying a Naïve Bayesian classification technique.

Chapter 7 introduces our approach to detect stops within a GPS track by clustering GPS points of a moving trajectory based on travel speed and distance. Our approach provides an advance over several previous methods by reducing the time cost of finding these clusters whilst identifying centroids of each identified cluster. The algorithm is also tested on user-labelled data that is collected for this research, which was an identified shortcoming in previous methods. In chapter 7, we also identify indoor activity using the dwell time concept, however separating underground travel from these instances using a validation process introduced in the following phase of the framework in chapter 8. Results from the classification and segmentation phases are then integrated and presented at the end of chapter 7, demonstrating the effect of combining the results of both processes together.

In **chapter 8**, we coin the term “Network Matching” as the process of selecting the network on which a trajectory is travelling. This phase verifies whether the modes identified from the SVM classification (developed in chapter 6) follow their respective transportation networks. The networks involved in this verification stage are restricted to public transit networks (bus, train and underground) since they are of a unique and single-mode nature. We also verify underground (tunnel) travel by applying a spatio-temporal reasoning algorithm using the London underground transport network to match to. The improvement in the accuracy of the mode of transport detection results after applying network matching is demonstrated at the end of chapter 8, highlighting the effect of the process on different modes.

The inference framework described in chapters 6, 7, and 8 are validated using the pilot data (21 users) collected initially for this research. In **chapter 9**, however, we test the framework using the main dataset collected for this research consisting of 95 participants. In chapter 9, also we also assess the effect of adding this prior knowledge on the performance of the inference framework developed in this thesis. The prior knowledge we use for this assessment is mainly participant-related information such as the ownership of a bike or a car, having a driving license, and access to local bike-rental services such as Barclays Bikes in London. We finally summarise outcomes of this study in **chapter 10**, discussing the conclusions, limitations, and lessons learnt from work done in this research. Chapter 10 also presents some proposed future work that could be conducted in order to extend on the work developed in this research.

Chapter 4

Data Collection

4 DATA COLLECTION⁴⁵⁶

Planning a positional data collection session depends entirely on the purpose of the collection. The chapter discusses four issues highlighted in chapter 2 in the data collection phase (as shown in Figure 4.1) for GPS-based travel surveys aiming to detect the mode of transport.

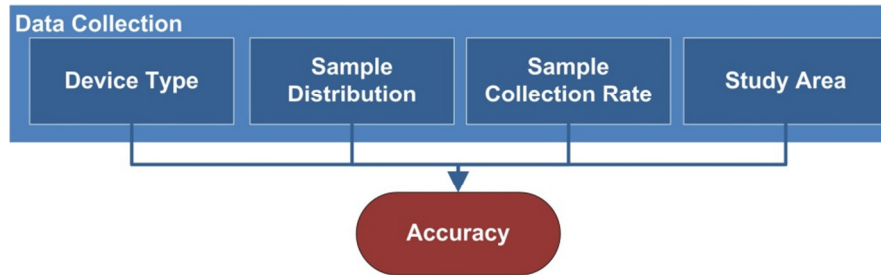


Figure 4.1 Data Collection Elements affecting Validation of Positional Survey Methods

This chapter starts by describing properties of existing devices used for positional data collection, and attempts to select the device with the best configuration to collect positional data which is best fit for detecting the mode of transport in Section 4.1. Section 4.2 discusses the basis of the selection process of the study area to conduct the intended study. Section 4.4 describes our attempt to calculate the minimum adequate sample participant size and study duration to collect GPS data for the study area of this research. In this context, we provide an understanding of the sample data collected for validate the method developed for inferring the mode of transport. This demonstration attempts to visualise and quantify the spatiotemporal distribution of the collected data in order to ensure its adequacy for testing the classification framework developed in this work. Finally, section 4.3 attempts to calculate the most appropriate rate of GPS data collection which minimises the battery consumption and the computational cost required for processing and analysing data from the chosen study area.

⁴ Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T and Tsapakis, I, 2014. A Spatio-Temporal Approach for Identifying the Sample Size for GPS-Based Travel Surveys: A Case Study of London's Road Network. *Journal of Transport Research: Part C*. Volume 43, Part 2, June 2014, Pages 176–187.

⁵ Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T, Tsapakis, I and Chow, A, 2012. Sample Size Calculation for Studying Transportation Modes from GPS Data, *Procedia - Social and Behavioral Sciences*, Volume 48, 2012, Pages 3040–3050.

⁶ Part of this chapter was presented in the following publication: Bolbol, A and Cheng, T, 2010. GPS Data Collection Setting For Pedestrian Activity Modelling. GISRUUK 2010: Proceedings of Geographical Information Science Research UK Conference, UCL, London, April 14–16, 2010.

4.1 Positioning Device Type

As described previously in chapter 2, most studies that have attempted to infer the mode of transport from positional data have collected the data using either: mobile applications on smart phones (Manzoni, et al., 2010; Stenneth, et al., 2011); strictly GPS devices alone (Chung & Shalaby, 2005; Liao, et al., 2007; Stopher, et al., 2008a; Schüssler & Axhausen, 2009; Zheng, et al., 2010) or integrated with other devices, such as accelerometers (Reddy, et al., 2010), or; others through mobile phone call detail records (CDR) (Wang, et al., 2010). This diversity of positioning sensor devices used for this type of studies is overwhelming in the capabilities that each encompasses. However, the selection of the most adequate device for the study must maintain the threshold between different properties such as the battery life, size, accuracy, coverage, interactivity and real-time capabilities. This could be noted from the selected devices for data collection in these previous studies, and is an important factor in the device selection process. This section gives a brief background of the available positioning sensor technologies nowadays. We also describe the properties of each of these technologies and discuss the most desirable properties modal detection studies require. We also introduce the selected devices to be used in this research, describing how they fulfil the needs of the collection circumstances they will be used in.

4.1.1 Positioning Systems

Location has proven to be one of the most important types of information for many of today's most critical mobile applications that are crucial to various industries (e.g. navigation, fleet management, etc.). For the purpose of this research, we need to collect data for tracking a human trajectory within different transport networks to detect the travel mode; therefore, this device used should provide a *wide coverage* and *good accuracy*.

One of the most commonly used positioning sensor devices nowadays is the **GPS**. The system uses triangulation from satellite positions in the sky through having a direct line of sight with each satellite. The GPS system is a U.S. space-based global navigation satellite system, and consists of 24 -32 satellites covering the earth, which means that, it covers almost every outdoor location on the earth's surface. The hand-held GPS devices could achieve an accuracy of 5m (Hinch, 2007). The disadvantages however are that it provides no indoor coverage, and sometimes there are some accuracy limitations due to some sources of error such as multi-paths, bad GDoP, etc.

GSM technology functions in the same way as GPS does, but it triangulates from mobile network base stations rather than satellites. The system provides a good coverage in urban areas and unlike GPS, it provides indoors coverage too (Schwieger, 2007). However, it provides a modest accuracy of 40m to 100m and does not provide coverage in non-GSM areas.

Wi-Fi Positioning System or **WPS** is one of the latest technologies to be used for positioning, and it depends on Wi-Fi coverage. Companies like Skyhook have built and maintain a large global database of Wi-Fi access points and their precise locations. This data may then be used by a mobile electronic device to triangulate a user's position (Gallagher, et al., 2009). One of the main advantages of WPS is that it provides a positioning solution indoors. However like GSM, it only provides coverage in certain areas that fall within the range of Wi-Fi signals, which probably exist in urban areas. The Wi-Fi hotspots database also must be constantly

updated to keep up with Wi-Fi hotspot changes. The WPS also provides an accuracy of 13m to 40m which is still worse than GPS in lots of cases (Stopher, 2008).

WPS might also be combined with mobile phone tower triangulation and GPS to provide reliable and accurate position data when GPS signals may be weak or intermittent under a wide range of conditions, including among tall buildings and indoors, (Gallagher, et al., 2009). Among the limitations of this combination is the battery constraint where these devices consume a huge amount of battery power to carry for more than one day.

RFID tags are commonly used nowadays for tracking animals and children since they are cheap, don't require a line of sight and have low battery consumption. They operate on radio waves, so they provide a high accuracy (Chon, et al., 2004); however, they provide coverage of 100m or slightly more, and therefore might not be useful for a human trajectory travelling in a road network.

| Technology | Outdoor | Indoor | Accuracy | Battery | Method |
|------------------------------------|--------------------------------|---------------|---------------|--------------------------|-----------------------------|
| GPS-based | • Everywhere | • N/A | $\leq \pm 5m$ | According to usage | Satellite triangulation |
| GSM Mobile telephone- based | • No coverage in non-GSM areas | • Available | 40m–100m | High consumption | Base stations triangulation |
| Wi-Fi | • Only Wi-Fi covered areas | • Available | 13m–40m | Low | Wi-Fi signals |
| RFID Tags | <i>Passive</i> | • Small range | Few cm | Non required | Radio waves |
| | <i>Active</i> | • Small range | Few cm | High consumption | |
| | <i>Battery Assisted</i> | • Large range | Few cm | Requires external source | |
| | <i>Passive</i> | | | | |

Table 4.1 Properties of different Sensor Technologies used for Positioning

Table 4.1 highlights some properties of the previously discussed positioning sensor device systems along with the method of operation of each of them. From this table, it could be noted that Wi-Fi and GSM technologies would not be ideal for tracking human trajectories for travel behaviour studies in non-urban areas due to signal coverage problems and for battery constraints. The RFID technology is inappropriate for tracking humans in a big city or country neither, due to transmission capacity. On the other hand, while GPS could have a huge disadvantage of not providing indoor coverage, but it could provide a useful means for tracking transport patterns, within different transport networks, and the absence of data might even indicate the type of transport used (e.g. underground).

4.1.2 Mobile Technology Functionality Modes

Positioning devices function in various modes according to the purpose which the data collection process is carried out for and the device used. According to Stopher (2008), mobile positioning technologies could be categorized according to 3 aspects, namely: portability, interactivity and processing mode. This categorization describes the different ways mobile devices are used, and helps show what type of device is appropriate for which purpose it is intended for.

Considering **portability**, mobile devices could either be In-Vehicle or Portable. **In-Vehicle** devices have the advantage of not requiring a power source of their own, while **Portable** devices could be carried in a user's pocket all day, and not restricted to a vehicle usage. For detection of the travel mode, this research will use a portable device to be able to track human trajectories during different travel modes.

Considering **interactivity**, mobile devices could be either Interactive or Passive. **Interactive** devices are devices that expect the user to enter metadata along with position through PDA devices, smart phones, etc. **Passive** devices on the other hand, just record position, and no entry is required. Users could later enter data as they wish as in the case of travel diaries. For this research; passive devices are used to take off the burden of data entry every step along the way, and to make the process an easier and automatic one.

Considering the **processing mode**, devices could be either Real-Time (online) or Logged Data (offline) (Yin & Wolfson, 2004). **Real-Time** devices are useful for studying human behaviour only if there is a need to communicate with the respondent to ask questions about a particular activity. On the other hand, **Logged data** devices seem to be used in almost all applications nowadays. For a second-by-second recording, a one day worth of data would equal 86400 data points, therefore, a minimum of 4 MB dedicated to recording position information would be adequate for effective logging. This research uses a Logged data device type for battery and user burden purposes.

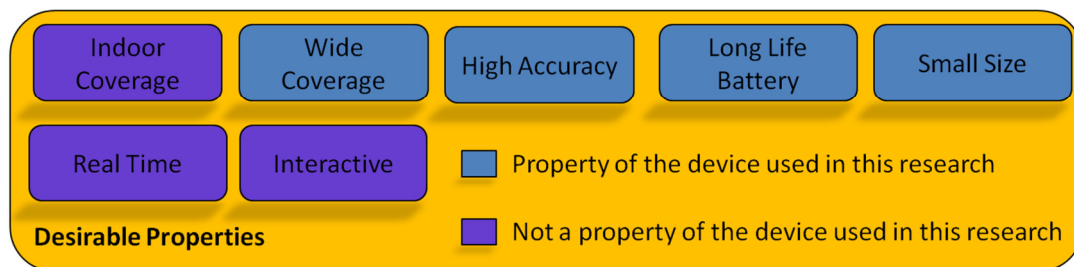


Figure 4.2 Desirable Properties of Positioning Devices

To sum up, among the desirable properties of positioning devices are that they provide indoor coverage, real time solution, wide coverage, high accuracy, long life battery and are interactive and small in size (as illustrated in Figure 4.2). However, for the purpose of this research we only require certain properties that fulfil specific requirements.

4.1.3 Mobile Devices for Mode of Transport Inference Studies

Hence, we use two types of devices in this research, both similar in concept and functionality. Both the selected devices fulfil the needs that this kind of research requires. Both the devices are based on GPS technology with a low battery consumption, high accuracy, non-real time, wide coverage and small in size.

The first device this research uses is a u-blox GPS logger. The logger is used for a pilot dataset collection to initially test the developed framework. The device is portable and small in size (5X3X1 cm) as illustrated on the left side of Figure 4.3 (u-blox, 2009). It also has a large battery capacity that lasts for almost a couple of months; to be able to track users in all travel modes

in an unbiased environment. It is also passive (non-interactive piece of hardware) for battery purposes and to keep the user's attention away from it to insure an unbiased behaviour. The processing mode is of an offline mode (post-processing) to take off the burden of data entry every step along the user's journey. The device is also equipped with an accelerometer that activates the data collection mode only when motion is detected. This also helps increasing the battery life massively. The device was developed by u-blox which contributes in funding and industrial supervision of this very research.



u-blox GPS Device used to collect the Pilot Dataset



Gtrek II GPS Trip recorder used to collect the Validation Dataset

Figure 4.3 GPS Tracker Devices used to collect Data for this Research

The second device is also a handheld GPS logger used for the validation dataset survey. The device is a Gtrek II GPS Trip Recorder (right of Figure 4.3). It is about the size of a cell phone, much more portable than a car-based GPS logger. The device has a very wide coverage and only fails to log when it is underground or in tunnels. Whenever aboveground, it records positions inside buildings, buses, elevated trains, bridges, ferries, and in urban canyons, although the positions recorded may not be always accurate. The logger was pre-set in both surveys to record variables such as date, time, latitude, longitude, speed, number of satellites used (NSAT), and horizontal dilution of precision (HDOP). NSAT is the number of satellites that a GPS logger used to calculate its position. The greater the NSAT value, the more accurate the calculation is likely to be. HDOP is an index to describe how well the positions of the satellites, used to calculation latitude and longitude, are arranged in the sky at the time of the recording. The greater the HDOP value, the less accurate the calculation is.

4.2 Study Area

Study areas used for testing mode of transport detection algorithms varied in different studies from one city to several cities. However, some studies test the developed algorithm in non-complex areas (Ren & Karimi, 2009). On the other hand, others do not test the algorithm on a large portion of the transport networks of the study city due to the limited amount of the data collected. This is attributed to the small sample size chosen for the validation of the performance of the method.

In this study, we select the City of London as the testing study area. The data is collected for participants that are based or work in London, meaning that they have daily trips within London. London is selected due to its complexity and the diversity of its transportation networks. These networks include the road, underground, train, footpath and bus networks. The diversity of modes of transport that use these networks makes the classification problem a very difficult one. An illustrative map of London is presented in Figure 4.4 with all the transport networks that it supports.

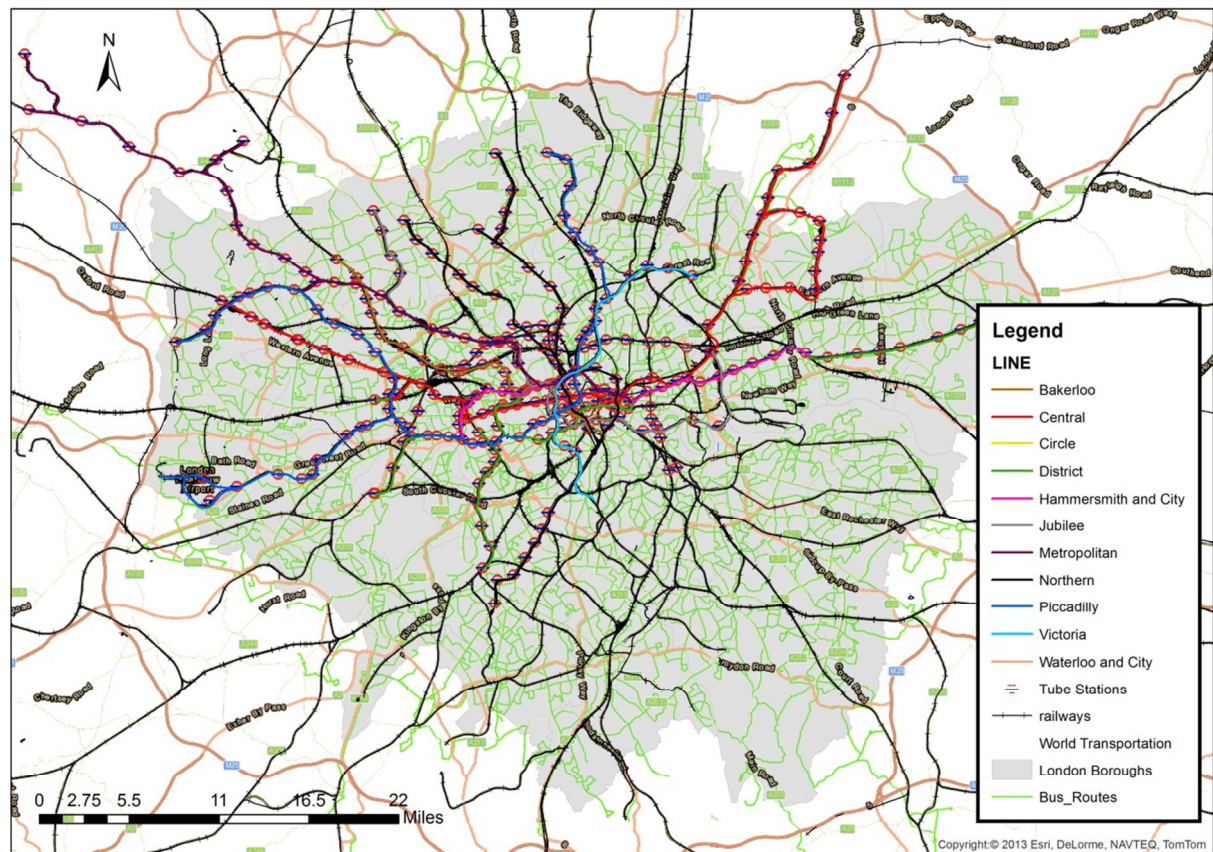


Figure 4.4 London & its Diverse Transport Networks

4.3 Sample Collection Rate

As described in chapter 2, one of main sampling problems is defining the GPS data collection rate (or the epoch rate), where a very fine rate of collection holds limitations. Although it might seem obvious that the more data the better, however, too much data will cause problems such as overloading the memory and battery power of current GPS devices or smart phones. Moreover, other problems include increasing the computation cost of the used

algorithms which gives rise to problems like increasing the burden on the participants to charge their devices very frequently and act as a constant reminder that they have a tracking device. Furthermore, the epoch rate chosen has a significant impact on the GPS data which results in GPS data errors such as “Cold Starts” (i.e. the device doesn’t begin recording at the exact start/end location of a trip) or “Map Mismatching” (i.e. problems of lateral movement of position from the GPS trace) (Stopher, 2008).

There are attempts to find the suitable epoch rate for monitoring animals (Perotto-Baldivieso, et al., 2008) however, not much research provides detailed analysis of the best GPS epoch rate setting for tracking participants within an urban environment. Moreover, all studies aiming at inferring the mode of transport from this data tend to use the finest grained temporal granularity for collecting their data without any statistical or logical justification. Therefore, this section attempts to find the most suitable epoch rate for an offline GPS device attempting to identify the mode of transport, as well as providing a statistical justification for the selection process. In broader terms, this study aims at providing the best configuration of a GPS device in order to have more meaningful data that arguably would lead to a better representation of human travel behaviours.

4.3.1 Data

We have collected data within the area of London to test our collection rate study using the GPS device produced by u-blox. This device has a positional accuracy of ± 4.3 meters (u-blox, 2009); however, the accuracy is significantly affected by GPS systematic sources of errors in urban areas. Three datasets were collected simultaneously for around a 7-8 minute walking journey, which is assumed to be a typical pedestrian walking journey (e.g. from car-work, home-car, home-bus stop, etc.). The tracks are illustrated in Figure 4.5. The walk mode was specifically chosen due to the complexity of a pedestrian trajectory when walking into narrow urban canyons.



Figure 4.5 1 Second Dataset & actual Route taken

Data was collected every 1 second within the area of the City of London, starting at a source location and ending at a final location. Three different carrying positions were attempted so as to vary the device handling technique to avoid biased outcome. The details of these datasets are demonstrated in the Table 4.2.

| Dataset | Mode | Base data's epoch rate | Position |
|---------|---------|------------------------|-------------|
| 1 | Walking | 30 Seconds | Hand |
| 2 | Walking | 1 Second | Back Pack |
| 3 | Walking | 1 Second | Back Pocket |

Table 4.2 Datasets with Different Holding Positions

The data was thinned in Microsoft Excel to the following epoch rates: 1, 10, 20, 30, 60, 120 and 300 seconds. The dataset was thinned 11 times in order to study the different probabilities from having 11 different datasets from every epoch rate group. From Figure 4.6, it could be noted that at different thinning levels some epoch rate sets could lose some data due to lack of fixes at particular instances.

| Dataset 1 - Thin 1 | | | | | | | | | Dataset 1 - Thin 2 | | | | | | | | | Dataset 1 - Thin 3 | | | | | | | | | Dataset 1 - Thin 4 | | | | | | |
|--------------------|-------|-------|-------|-------|-----|-----|----|-----|--------------------|-------|-------|-------|-------|-----|-----|-----|-------|--------------------|------|------|------|-----|-----|-----|-------|-------|--------------------|-------|-------|-----|-----|--|--|
| 1 | 10 | 20 | 30 | 60 | 120 | 300 | T | T+1 | 1 | 10 | 20 | 30 | 60 | 120 | 300 | T+2 | 1 | 10 | 20 | 30 | 60 | 120 | 300 | T+3 | 1 | 10 | 20 | 30 | 60 | 120 | 300 | | |
| 19.13 | | | | | | | 55 | 56 | 19.13 | | | | | | | 57 | 19.13 | | | | | | | 58 | 19.13 | | | | | | | | |
| 17.34 | | | | | | | 56 | 57 | 17.34 | | | | | | | 58 | 17.34 | | | | | | | 59 | 17.34 | | | | | | | | |
| 17.15 | | | | | | | 57 | 58 | 17.15 | | | | | | | 59 | 17.15 | | | | | | | 60 | 17.15 | 17.15 | 17.15 | 17.15 | 17.15 | | | | |
| 6.88 | | | | | | | 58 | 59 | 6.88 | | | | | | | 60 | 6.88 | 6.88 | 6.88 | 6.88 | 6.88 | | | 61 | 6.88 | | | | | | | | |
| 18.95 | | | | | | | 59 | 60 | 18.95 | 18.95 | 18.95 | 18.95 | 18.95 | | | 61 | 18.95 | | | | | | | 62 | 18.95 | | | | | | | | |
| 28.85 | 28.85 | 28.85 | 28.85 | 28.85 | | | 60 | 61 | 28.85 | | | | | | | 62 | 28.85 | | | | | | | 63 | 28.85 | | | | | | | | |
| 11.95 | | | | | | | 61 | 62 | 11.95 | | | | | | | 63 | 11.95 | | | | | | | 64 | 11.95 | | | | | | | | |
| 11.78 | | | | | | | 62 | 63 | 11.78 | | | | | | | 64 | 11.78 | | | | | | | 65 | 11.78 | | | | | | | | |
| 13.15 | | | | | | | 63 | 64 | 13.15 | | | | | | | 65 | 13.15 | | | | | | | 66 | 13.15 | | | | | | | | |
| 5.31 | | | | | | | 64 | 65 | 5.31 | | | | | | | 66 | 5.31 | | | | | | | 67 | 5.31 | | | | | | | | |
| 12.33 | | | | | | | 65 | 66 | 12.33 | | | | | | | 67 | 12.33 | | | | | | | 68 | 12.33 | | | | | | | | |
| 24.69 | | | | | | | 66 | 67 | 24.69 | | | | | | | 68 | 24.69 | | | | | | | 69 | 24.69 | | | | | | | | |
| 29.00 | | | | | | | 67 | 68 | 29.00 | | | | | | | 69 | 29.00 | | | | | | | 70 | 29.00 | 29.00 | | | | | | | |
| 21.31 | | | | | | | 68 | 69 | 21.31 | | | | | | | 70 | 21.31 | 21.31 | | | | | | 71 | 21.31 | | | | | | | | |
| 32.38 | 32.38 | | | | | | 70 | 71 | 32.38 | | | | | | | 72 | 32.38 | | | | | | | 73 | 32.38 | | | | | | | | |
| 23.13 | | | | | | | 71 | 72 | 23.13 | | | | | | | 73 | 23.13 | | | | | | | 74 | 23.13 | | | | | | | | |
| 22.54 | | | | | | | 73 | 74 | 22.54 | | | | | | | 75 | 22.54 | | | | | | | 76 | 22.54 | | | | | | | | |
| 6.75 | | | | | | | 81 | 82 | 6.75 | | | | | | | 83 | 6.75 | | | | | | | 84 | 6.75 | | | | | | | | |
| 11.43 | | | | | | | 83 | 84 | 11.43 | | | | | | | 85 | 11.43 | | | | | | | 86 | 11.43 | | | | | | | | |
| 11.84 | | | | | | | 85 | 86 | 11.84 | | | | | | | 87 | 11.84 | | | | | | | 88 | 11.84 | | | | | | | | |
| 24.93 | | | | | | | 87 | 88 | 24.93 | | | | | | | 89 | 24.93 | | | | | | | 90 | 24.93 | 24.93 | | 24.93 | | | | | |
| 0.42 | | | | | | | 88 | 89 | 0.42 | | | | | | | 90 | 0.42 | 0.42 | | 0.42 | | | | 91 | 0.42 | | | | | | | | |
| 3.49 | | | | | | | 89 | 90 | 3.49 | 3.49 | | 3.49 | | | | 91 | 3.49 | | | | | | | 92 | 3.49 | | | | | | | | |
| 10.41 | 10.41 | | | 10.41 | | | 90 | 91 | 10.41 | | | | | | | 92 | 10.41 | | | | | | | 93 | 10.41 | | | | | | | | |

Figure 4.6 Missing Data due to Lack of Fixes in a Dataset

4.3.2 Testing Methodology

In the context of this research, we aim to collect GPS data for inferring travel details. Therefore, we consider that attributes such as detecting the route and mode of transport are the most significant factors affecting such an inference model. As illustrated in Figure 4.7, four accuracy indicators are tested by comparing four corresponding accuracy measures namely; positional errors, route length errors, average speed errors and distances from trip end points to the origin and the destination locations.

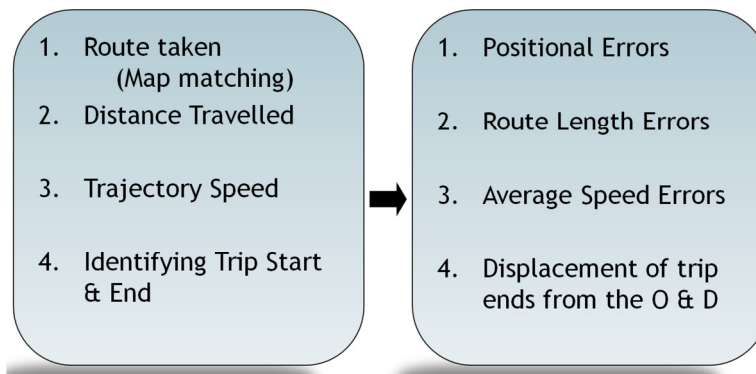


Figure 4.7 Attributes that form the Epoch Rate Comparison Test

4.3.2.1 Calculating Positional Error

The positional error is probably the most important factor to be tested. This is mainly due to different sources of GPS systematic errors such as the Ionosphere and Troposphere disturbances, signal reflection, ephemeris errors, clock errors, visibility of satellites, satellite shading and defined cut off angle (Hinch, 2007). Other sources of errors are due to specifications of the u-blox GPS device used. The effect of bad positional accuracy could result in many problems such as the following:

- Error in associating the track to the correct road taken (map matching), in a scenario where roads are parallel and very close to each other
- Computing inaccurate distances and speeds values
- Effect of wandering and long stop errors (described in chapter 2)
- Inaccurate Identification of a trip's origin , stops and destination

An algorithm was developed and coded in the R Project for Statistical Computing (R Project, 2012) that measures the positional displacement (positional accuracy) of GPS records (points) from the actual path/route (polyline) taken by the user while collecting the data. The algorithm finds the shortest distance from a given GPS fix to the nearest road link. The algorithm calculates the perpendicular distance from a GPS fix to the nearest road link and to each of the nodes of that road link. The algorithm then selects the shortest distance out of the three values for every GPS record to the road network (Figure 4.8).

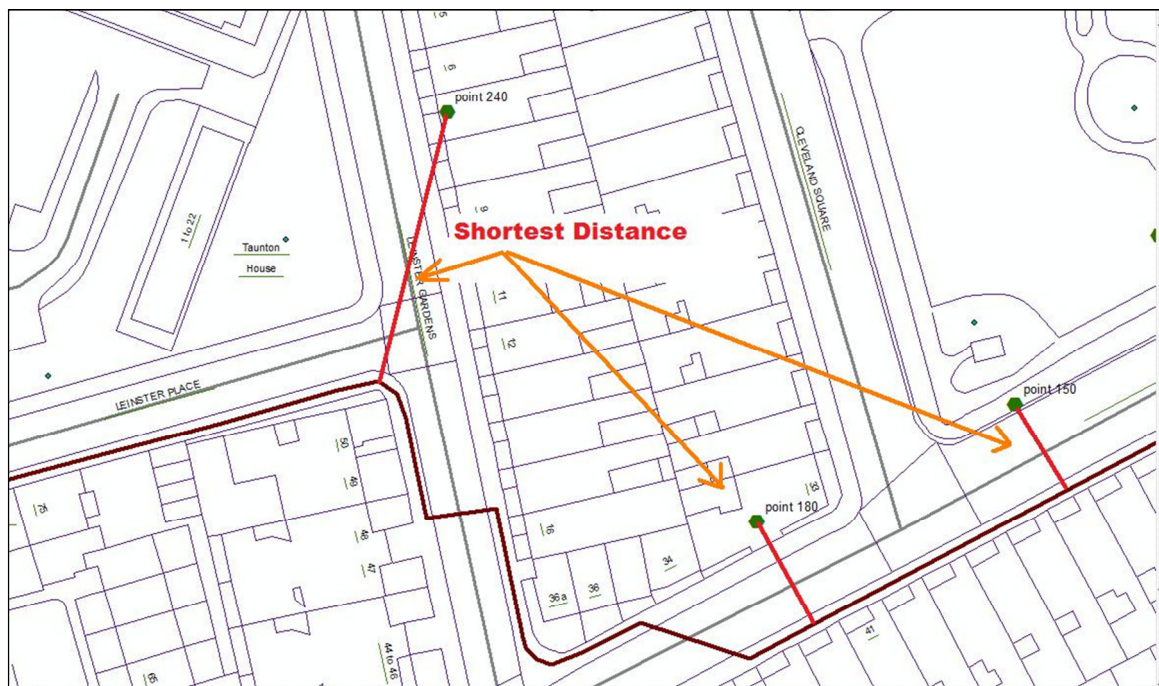


Figure 4.8 Shortest Distance calculation Method

4.3.2.2 Calculating Route Length Errors

Euclidean distances were calculated between every two successive GPS points for each data set, along with the speed. We did that by extracting the Longitude and Latitude (ϕ , λ)

coordinates to transforming them into Northing and Easting (N, E) coordinates going through a datum transformation (Airy1830) and a UTM projection into the National British Grid coordinates. The annotation below expresses this process. The total length of the track is later calculated from each dataset and compared with the actual length of the track.

$$(\phi, \lambda) \rightarrow (\text{Airy1830}) \text{ UTM} \rightarrow (N, E) \rightarrow (\Delta N, \Delta E) \rightarrow \text{Dist} \rightarrow \Sigma \text{ Dist} \rightarrow \Delta \Sigma \text{ Dist} \quad \text{Equation 4.1}$$

This will hence show the difference between different sessions and the actual route taken, and will express a value for that difference in distance calculation accuracy.

4.3.2.3 Calculating Average Speed Errors

By the knowledge of the distances and time interval between each two records, the speed is then determined for each GPS segment, and the average speed along the whole route is calculated along with the standard deviation (δ). The annotation below explains the process.

$$\text{Dist} + \text{Time Interval/Epoch Rate} \rightarrow \text{Speed} \rightarrow \Sigma \text{ Speed} \rightarrow \Delta \Sigma \text{ Speed} + (\delta) \quad \text{Equation 4.2}$$

4.3.2.4 Error in Trip's Origin and Destination

The aim of this test is to identify the origin and destination locations from where the GPS tracks start/end. This depends on the last points in the dataset. Given that there are 11 datasets thinned from one, the last point will always change except for the 1s dataset. The distance between the last point from every dataset and the end destination is measured (Figure 4.9).



Figure 4.9 Route with Data from different Datasets near End Destination

4.3.2.5 Map Matching Error

Map matching is first applied to all the thinned datasets. As described in chapter 2, map matching is a method of snapping the GPS points to the road link it was on. For a vehicle trajectory, topological map matching approaches seem to reveal superior performance compared with other existing algorithms (Quddus, 2006). Topological approaches usually perform better due to the fact that vehicles have to follow road restrictions such as one-way directions, U-turns, etc. However, in a pedestrian-based trajectory, topological rules might not apply. Therefore, a more suitable map matching technique in the case of this study would be a simple algorithm using a point-to-curve approach together with a heading and speed taken into consideration.

In order to quantify the map matching error, we match each fix to the nearest road link, then we use a similar method used in subsection 4.3.2.1 to measure positional errors for each GPS fixes. An example is illustrated below in Figure 4.10 for a 30 second session from dataset 2 showing the selected road links highlighted in blue for a group of fixes in a GPS track.

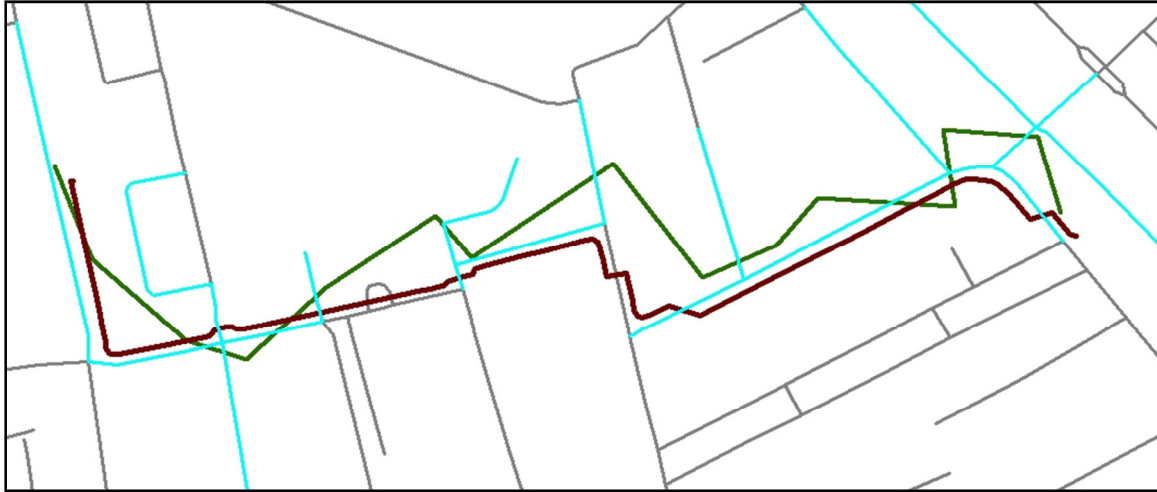


Figure 4.10 Map Matching selected Road Links for 30 seconds Track- Dataset 2

There are 14 road links being used in the test route. A buffer is run around each GPS session trail to the nearest road and a number of roads are selected due to each session. This number consists of some roads that are part of the actual track and some which are (false) roads. In order to quantify the efficiency of the buffer selection of each session, a couple of accuracy measures were defined as follows:

$$\text{Efficiency of Selection} = \frac{\text{Number of Right Selected Roads}}{\text{Total Number of Selected Roads}} \quad \text{Equation 4.3}$$

$$\text{Efficiency of Right Selection} = \frac{\text{Number of Right Selected Roads}}{\text{Total Number of Roads involved in Actual Route}} \quad \text{Equation 4.4}$$

And finally, in order to calculate an average percentage:

$$\text{Accumulated Efficiency} = \frac{\text{Efficiency of Selection} + \text{Efficiency of Right Selection}}{2} \quad \text{Equation 4.5}$$

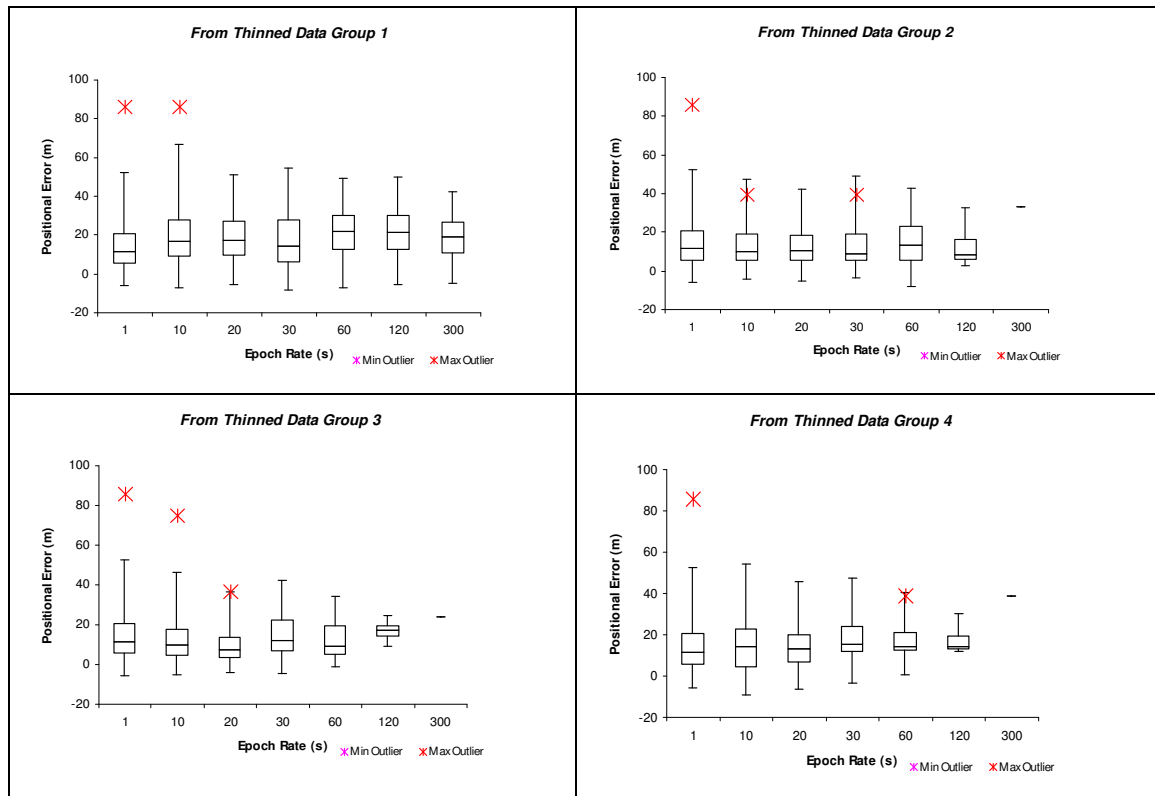
4 Data Collection

4.3.3 Results and Analysis

The results of this test help understand the impact of using different epoch rates of GPS data collection on a trajectory representation. It also provides us with recommendations for the most appropriate epoch rates for different tasks. This section illustrates and discusses the results of this test, and provides the guidelines for choosing appropriate epoch rates for different purposes.

4.3.3.1 Positional Error

Only 55 records out of 470 did not get a fix, and 2 out of 416 points had accuracy worse than 60m. Figure 4.11 shows box-plots illustrating the different positional errors of each of the 11 thinned datasets with a combined summary plot at the end of the figure. The averages of all datasets tend to be close in value; giving an accuracy of 15-20 meters. There aren't many points in the 300 second datasets and sometimes none; therefore, they prove not to be sufficient for data collection in that context. The 1 second dataset on the other hand has more variances and instances which might badly affect the map matching process. The following are the results from the 11 thinned datasets presented in box plots to visualise the impact of using longer epoch rates on the positional error and hence; reducing the number of fixes used.



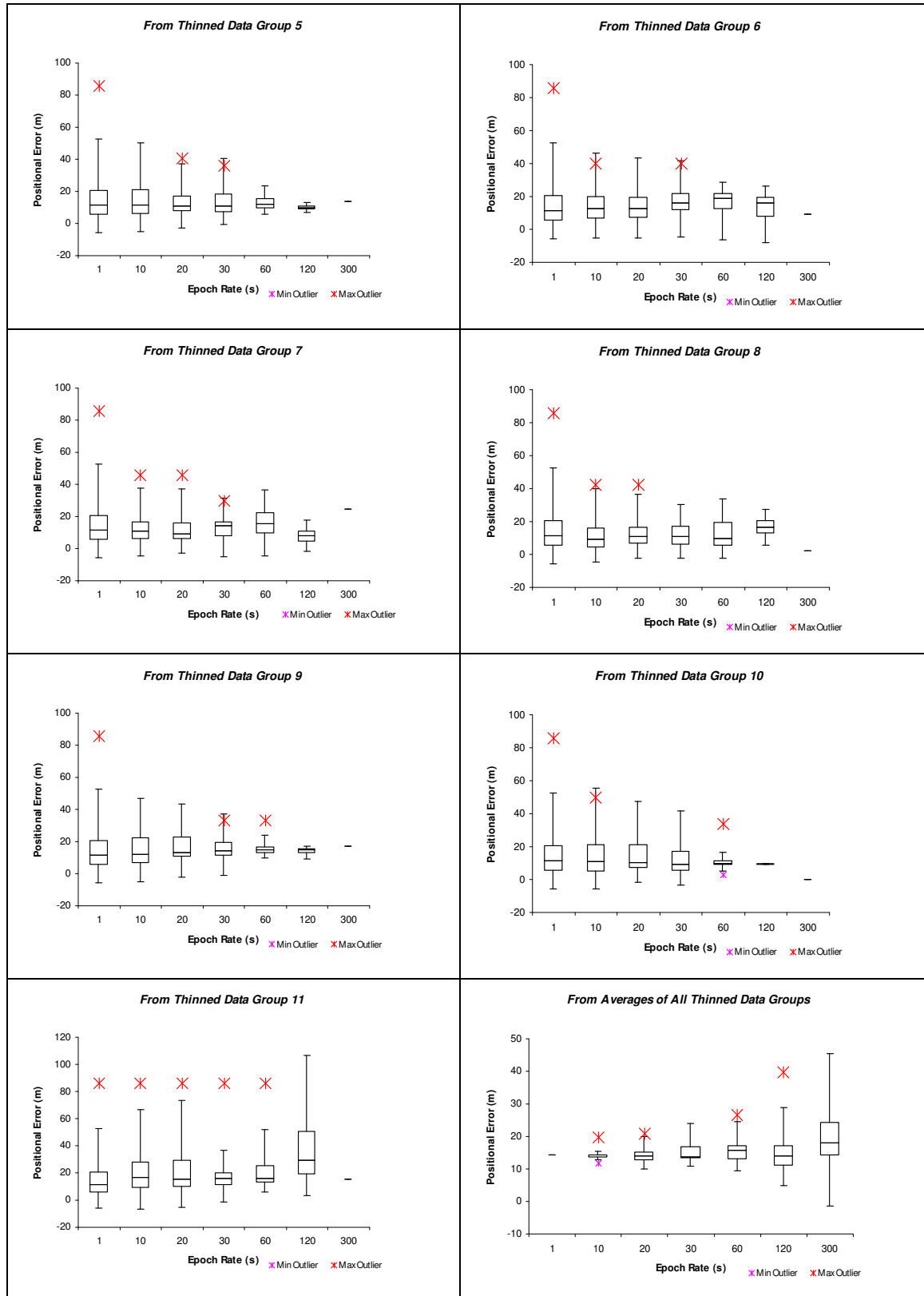


Figure 4.11 Box-Plots illustrating Positional Error of the 11 thinned Datasets

4.3.3.2 Route Length Error

Results of the length and speed are shown in Table 4.3 and as a box plot representation in Figure 4.12. The lengths of the route computed from 20 second data prove to be the closest to

the actual route length followed by the 10 and 30 second data, however, there is a higher certainty with the 30 second sessions due to closeness of their results. 1 second results have been discarded from Figure 4.12 due to their high error values, while only one thinned set was extracted for 300 second data (marked as x in Table 4.3).

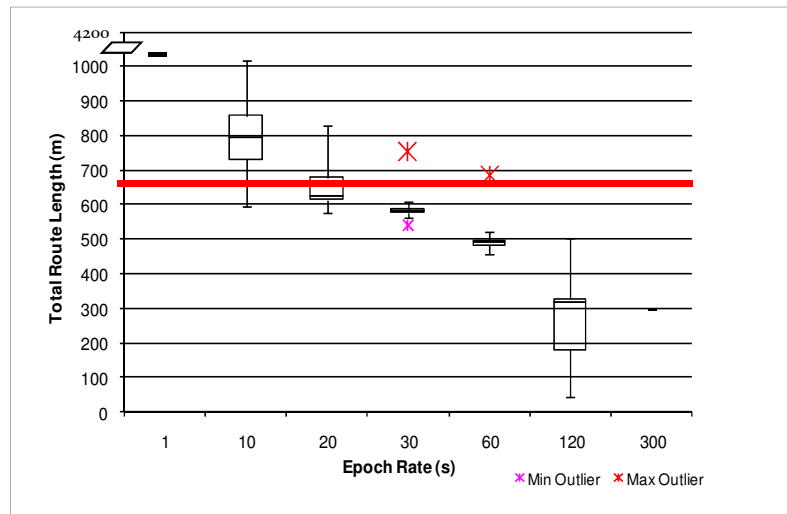


Figure 4.12 Total Route Length calculated from Datasets Compared to Actual Length

| 1 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---------|-------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Length | 4143.03 | | | | | | | | | | |
| Speed | 9.33 | | | | | | | | | | |
| St Dev | | | | | | | | | | | |
| 10 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 952.52 | 766.0 | 834.7 | 829.0 | 739.2 | 690.0 | 659.7 | 723.4 | 793.9 | 879.5 | 952.5 |
| Speed | 2.16 | 1.72 | 1.93 | 1.78 | 1.49 | 1.50 | 1.50 | 1.61 | 1.90 | 1.98 | 2.16 |
| St dev | 2.25 | 1.78 | 2.02 | 1.77 | 1.47 | 1.58 | 1.80 | 1.66 | 1.86 | 1.69 | 2.25 |
| 20 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 643.32 | 680.9 | 595.5 | 746.8 | 585.1 | 615.0 | 619.0 | 611.5 | 679.1 | 622.6 | 773.6 |
| Speed | 1.42 | 1.59 | 1.37 | 1.70 | 1.28 | 1.44 | 1.41 | 1.47 | 1.59 | 1.38 | 1.63 |
| St dev | 0.93 | 1.05 | 0.81 | 1.21 | 0.92 | 1.02 | 0.88 | 0.88 | 0.74 | 0.92 | 1.58 |
| 30 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 600.31 | 580.5 | 585.9 | 580.4 | 555.0 | 580.0 | 576.5 | 587.9 | 538.8 | 586.8 | 754.1 |
| Speed | 1.35 | 1.43 | 1.42 | 1.37 | 1.34 | 1.40 | 1.36 | 1.40 | 1.32 | 1.33 | 1.72 |
| St dev | 0.70 | 0.57 | 0.48 | 0.61 | 0.55 | 0.49 | 0.52 | 0.56 | 0.78 | 0.56 | 1.04 |
| 60 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 543.57 | 466.8 | 484.1 | 484.1 | 495.7 | 493.8 | 500.4 | 492.4 | 480.8 | 493.6 | 683.0 |
| Speed | 1.29 | 1.33 | 1.37 | 1.34 | 1.38 | 1.37 | 1.37 | 1.37 | 1.34 | 1.33 | 1.63 |
| St dev | 0.38 | 0.36 | 0.33 | 0.34 | 0.31 | 0.25 | 0.27 | 0.23 | 0.26 | 0.15 | 0.63 |
| 120 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 422.33 | 315.7 | 178.3 | 315.7 | 315.7 | 313.6 | 180.5 | 178.3 | 339.6 | 180.5 | 490.9 |
| Speed | 1.17 | 1.32 | 1.49 | 1.32 | 1.32 | 1.31 | 1.50 | 1.49 | 1.42 | 1.50 | 1.36 |
| St dev | 0.29 | 0.17 | 0.00 | 0.17 | 0.17 | 0.20 | 0.00 | 0.00 | 0.07 | 0.00 | 0.28 |
| 300 Sec | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Length | 298.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Speed | 0.994 | | | | | | | | | | |
| St dev | 1 | | | | | | | | | | |
| Actual | Actual Route Data | | | | | | | | | | |
| Length | 667.00 | | | | | | | | | | |
| Speed | 1.20 | | | | | | | | | | |
| | Length (m), Speed (m/s) | | | | | | | | | | |

Table 4.3 Route Length & Average Speed from the 11 Datasets & Actual Route

4.3.3.3 Average Speed Errors

The average speed results shown in Figure 4.13 demonstrate that 30 and 60 second data give the best results and have a high level of certainty, followed closely by the 120 and 20 second datasets. 1 second results were again discarded due to their high error value, while the 300 second data has only one result because only one 30 second thinned-dataset had 2 points.

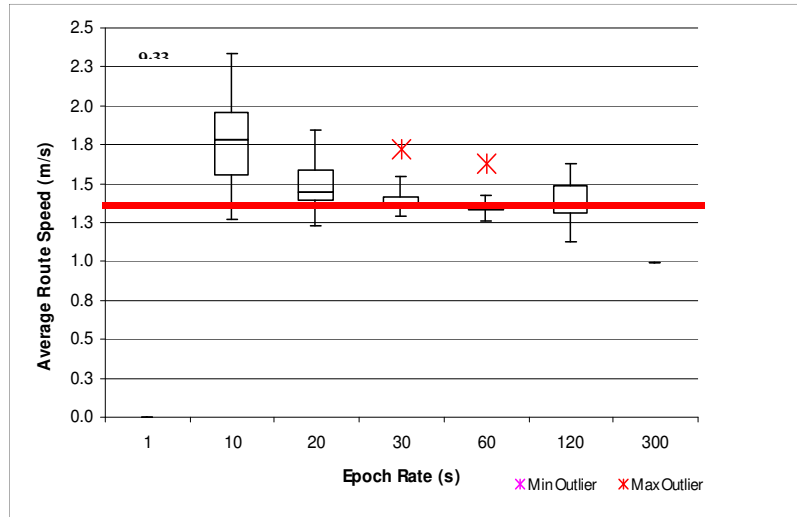


Figure 4.13 Average Speed calculated from different Datasets compared to Truth Average Speed

4.3.3.4 Distance to Trip Ends

Since the 1 second data is only 1 Dataset (with no thinned versions), the trip ends results are the same for all datasets, because only one point is chosen in all cases of data thinning. However, in any probabilistic model aiming at identifying the start/end location of a trip, the excessive data obtained from 1 second epoch rate makes it more likely to get better results but with high uncertainty. Nevertheless, results from 10 and 20 second data from Figure 4.14 and Table 4.4 seem sufficient enough to infer the start s and end locations. This is supported by the fact that GPS accuracy within an urban area is around 15-20 meters (section 4.3.3.1), and adding to that the length it takes a 10 or 20 second device to get a fix is around 12 or 24 meters respectively (assuming speed=1.38 m/s); would result in a distance of maximum 45 meters under the worst case scenario with 20 seconds.

| | 1s | 10s | 20s | 30s | 60s | 120s | 300s |
|-----------|-----|-----|-----|------|------|-------|-------|
| Dataset1 | 1.6 | 1.6 | 1.6 | 1.6 | 45.3 | 116.8 | 210.0 |
| Dataset2 | 1.6 | 1.6 | 1.6 | 1.6 | 45.3 | 120.7 | 210.0 |
| Dataset3 | 1.6 | 8.0 | 1.6 | 8.0 | 45.3 | 120.7 | 210.0 |
| Dataset4 | 1.6 | 0.1 | 1.6 | 1.6 | 45.3 | 120.7 | 210.0 |
| Dataset5 | 1.6 | 8.0 | 8.0 | 1.6 | 45.3 | 120.7 | 198.3 |
| Dataset6 | 1.6 | 1.6 | 0.1 | 1.6 | 45.3 | 120.7 | 198.3 |
| Dataset7 | 1.6 | 1.6 | 1.6 | 8.0 | 45.3 | 120.7 | 187.3 |
| Dataset8 | 1.6 | 1.6 | 1.6 | 8.0 | 66.4 | 120.7 | 203.6 |
| Dataset9 | 1.6 | 1.6 | 1.6 | 66.4 | 66.4 | 120.7 | 196.3 |
| Dataset10 | 1.6 | 1.6 | 0.1 | 8.0 | 66.4 | 120.7 | N/A |
| Dataset11 | 1.6 | 1.6 | 1.6 | 1.6 | 1.6 | 1.6 | 193.6 |

Table 4.4 Start-End Trip Distance Results (m)

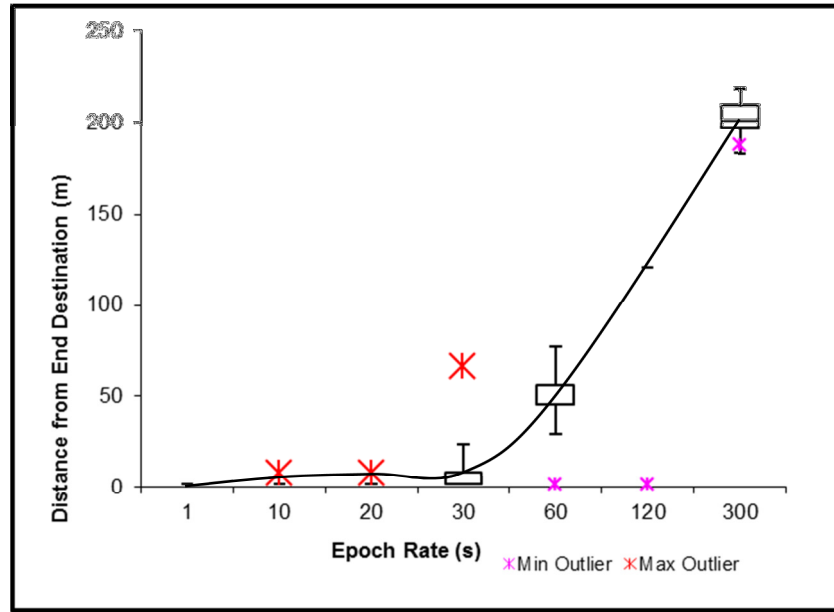


Figure 4.14 Distances from Set of Last Points in Each Dataset to End Destination

4.3.4 Experiment Conclusion

The study has shown that u-blox GPS devices achieve a positional accuracy of around 15-20 meters within an urban area. In long epoch rate sessions, some bad quality data might influence the whole dataset to have a bad overall positional quality. On the other hand, a 1 second dataset might lead to bad map matching due to excessive scattered data nature.

The study has also shown that 20 second datasets prove to be superior for route length calculation, while that 120 second data achieves low accuracy. It also shows that 60 second data gives the best accuracy for calculating a pedestrian's speed. As for determining the trip ends, the 10 and 20 second data achieve the best results, bearing in mind that 1 second data also achieves high accuracy but with high uncertainty.

To select the correct road links of every trip for the map matching process, we assume that we need at least a GPS point on each link of the road network to obtain a good probability of an accurate link choice. There are around 300,000 road links in London, and their average length is around 80 meters with a standard deviation of 98.47 meters. Assuming that the average speed of pedestrians = 1.38 m/s (Knoblauch, et al., 1996); this would mount to a suitable epoch rate of around 110 seconds. Nevertheless, a more conservative measure would be to use the fifth percentile (0.05) of road lengths (≈ 10 meters) along with the 95th percentile of speeds of all transport modes in the validation dataset collected for this research (≈ 2.8 m/s); mounting to a suitable collection rate of ≈ 30 seconds.

And as a whole, the study concludes that the most appropriate epoch rate for route and the trip ends detection of pedestrians is probably somewhere between 30 and 60 seconds. This could arguably increase the accuracy of mode of transport detection from GPS data. As mentioned earlier in this section, the walk mode is a useful mode to test GPS errors due to the complexity of nature of the movement of pedestrian trajectories that does not strictly follow any network rules or restrictions and that tends to usually move in narrow urban canyons. Nevertheless, more experiments could be carried out for other transport modes.

4.4 Sample Spatial, Modal and Temporal Distributions

As mentioned in chapter 2, sampling issues also include that most studies do not provide any understanding of the details for their sampling coverage either in the spatial or temporal domains. Most of these studies neither accounts for the weekly seasonal variation which creates a bias towards that specific week of collection. Moreover, none of the studies which provide a sample size and duration calculation framework for GPS-based travel surveys attempt to provide an understanding of the spatial, modal or temporal granularities of the designed study. Such comprehension of the spatial, modal and temporal extents underlines the context limitations of these studies. This section describes our attempt to clarify the extent of the collected data by providing an understanding of the sample spatial, modal and temporal distributions and duration of this study as an example of the sample assessment process for studies that aim at detecting the mode of transport from GPS data. We provide statistical analysis and plots of the distribution of the data within the spatial, modal and temporal domains in an attempt to highlight limitations of the data coverage to be noted when results of this thesis are taken for any further application or analysis.

4.4.1 Collected Datasets

In this research, two datasets are collected. First, a **pilot dataset** is collected consisting of 21 persons over 2 weeks using u-blox GPS devices (u-blox, 2009) where the mode of transport is labelled by the participant segment-by-segment. Second, a **validation dataset** is collected consisting of 95 people over 2 months using GTrek GPS devices (GTrek, 2012) and information of transport modes used and ownership of a bike, access to Barclays Bikes, driving license and ownership of a car is provided by the user for validation purposes.

All through the chapters of this thesis, the accuracy of different algorithms and processes are preliminarily evaluated using the pilot data (21 participants), whereas in chapter 9 (Further Validation), we apply all the algorithms to the validation dataset (95 participants – 11 drop-outs) to further evaluate the performance of the algorithms.

4.4.2 GPS Data Sample Spatial, Modal and Temporal Distributions in the Study Area

In this section, we attempt to understand and visualise the distribution of the GPS validation dataset (to be used in chapter 9 for assessing the framework's performance) in the spatio-temporal-modal domains. This section therefore allows us to note areas/times/modes where the validation dataset is not fully representative, and as a result, data limitations could be taken into consideration while validating the algorithm performance.

4.4.2.1 General Spatial Distribution

The validation dataset is collected mainly for people resident in London. The dataset however extends outside of the London area due to different travel activities. Figure 4.15 shows the distribution of the dataset within the area of the City of London divided TfL-defined traffic zones. The zones are doughnut-shaped dividing London into Central, Inner and Outer London, and the division is based on TfL's London Congestion Analysis Project (LCAP). The figure shows how the data is well distributed within areas of Central and Inner London. In Outer London however, the data seems to be sparser and slightly more focused on the areas of North and West London.

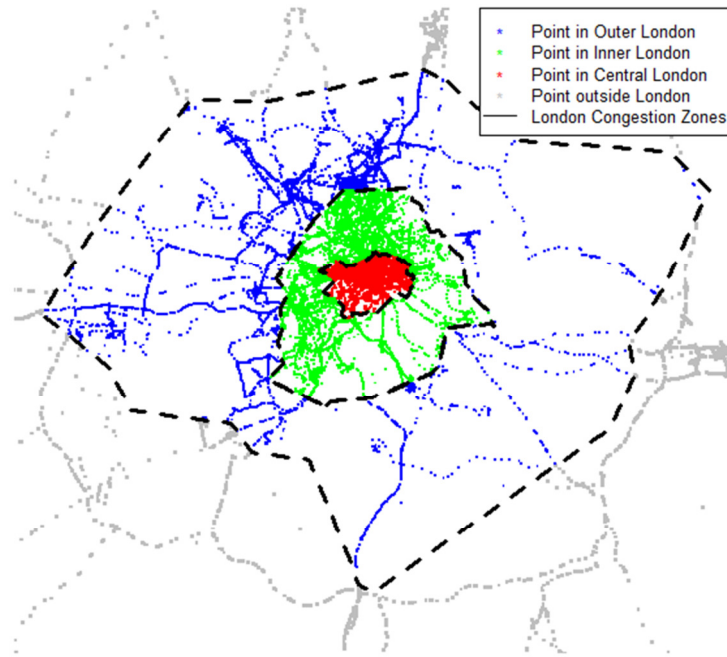


Figure 4.15 General Distribution of GPS Validation Dataset in London

Using MASS⁷ library in R Statistical Package, heatmaps were created for data in different London doughnut zones. Figure 4.16 shows these heatmaps for Central, Inner and Outer London zones from left to right respectively. Central London appears to have a high density in areas of central rail stations such as Euston, Oxford Circus, Paddington, King's Cross and a few at Victoria Train Station. There also seems to be a high density around University College London (UCL) premises where a portion of the participants work/study. Inner and Outer London seem to be spread around with the exception of areas that seem to act as hubs such as Hammersmith underground station with the rest of the data connecting between them along different transportation routes.



Figure 4.16 Validation Dataset Heatmaps of different Zones of London

⁷ <http://cran.r-project.org/web/packages/MASS/index.html>

4.4.2.2 Spatio-Temporal Distribution

Another means to assess the efficiency of distribution is by understanding the temporal distribution across different parts of the study area. Hence, the dataset is divided into significant periods of the day to represent different congestion peaks and troughs of the day as defined by TfL.

Table 4.5 provides a statistical distribution of the dataset in the spatio-temporal domain. As can be noted from the table, the night period appears to hold a big chunk of the data (35%). This is could be attributed to the active London night live and long periods when the participants are home to spend the night with GPS devices having coverage and continuously collecting data.

| Periods | Time | Central | Inner | Outer | Outside | Total |
|----------|-------------|---------|--------|--------|---------|--------|
| AM | 07:00-09:55 | 4.00% | 4.47% | 3.35% | 1.42% | 13.24% |
| Inter-AM | 10:00-12:55 | 4.63% | 2.75% | 2.57% | 1.57% | 11.52% |
| Inter-PM | 13:00-15:55 | 4.44% | 2.74% | 2.48% | 1.71% | 11.37% |
| PM | 16:00-18:55 | 3.66% | 3.97% | 2.76% | 1.65% | 12.04% |
| Evening | 19:00-21:55 | 2.72% | 4.61% | 3.41% | 1.78% | 12.52% |
| Night | 22:00-05:55 | 7.08% | 12.93% | 10.26% | 4.64% | 34.91% |
| Pre-AM | 06:00-06:55 | 0.95% | 1.63% | 1.30% | 0.53% | 4.41% |

Table 4.5 Spatio-Temporal Data Distribution in London (Colour Intensity-Coded according to Distribution Share)

Table 4.6, on the other hand, excludes the night period from the analysis in order to give a better understanding of the day-time spatio-temporal data distribution. The table shows the Central London to contain the highest activity during Inter-AM and Inter-PM periods (i.e. 10:00 to 16:00). Conversely, Inner and Outer London areas seem to hold the highest activity during AM, PM and evening periods. This reflects more outdoor movements during working hours in Central London and elsewhere in London outside of working hours, stressing the phenomena of the commute into Central London where many businesses are located. It could also be noted is the good spatio-temporal distribution (average 5.5%) of the dataset between different zones of London (excluding Pre-AM period which lasts only for an hour). The temporal distribution is also relatively uniform (average 18.5%) across different periods of the day.

| Periods | Time | Central | Inner | Outer | Outside | Total |
|----------|-------------|---------|-------|-------|---------|--------|
| AM | 07:00-09:55 | 6.14% | 6.87% | 5.15% | 2.18% | 20.34% |
| Inter-AM | 10:00-12:55 | 7.11% | 4.22% | 3.94% | 2.41% | 17.68% |
| Inter-PM | 13:00-15:55 | 6.82% | 4.21% | 3.81% | 2.63% | 17.47% |
| PM | 16:00-18:55 | 5.63% | 6.09% | 4.25% | 2.54% | 18.51% |
| Evening | 19:00-21:55 | 4.18% | 7.08% | 5.24% | 2.74% | 19.24% |
| Pre-AM | 06:00-06:55 | 1.47% | 2.50% | 2.00% | 0.81% | 6.78% |

Table 4.6 Spatio-Temporal Data Distribution in London Excluding Night Time Period (Colour Intensity-Coded according to Distribution Share)

Figure 4.17 shows this spatio-temporal distribution in the form of plots for every TfL time period of the day across the three congestion zones. The plots illustrate the fact that the night period contains the most movements as previously highlighted in Table 4.5. The plots also highlight the previous finding from Table 4.6 where lots of the travelling to Central London occurs during the AM and PM periods. It also is clear from the figure that movements appear to be even across different parts of Central and Inner London for most periods while variant in parts of Outer London.

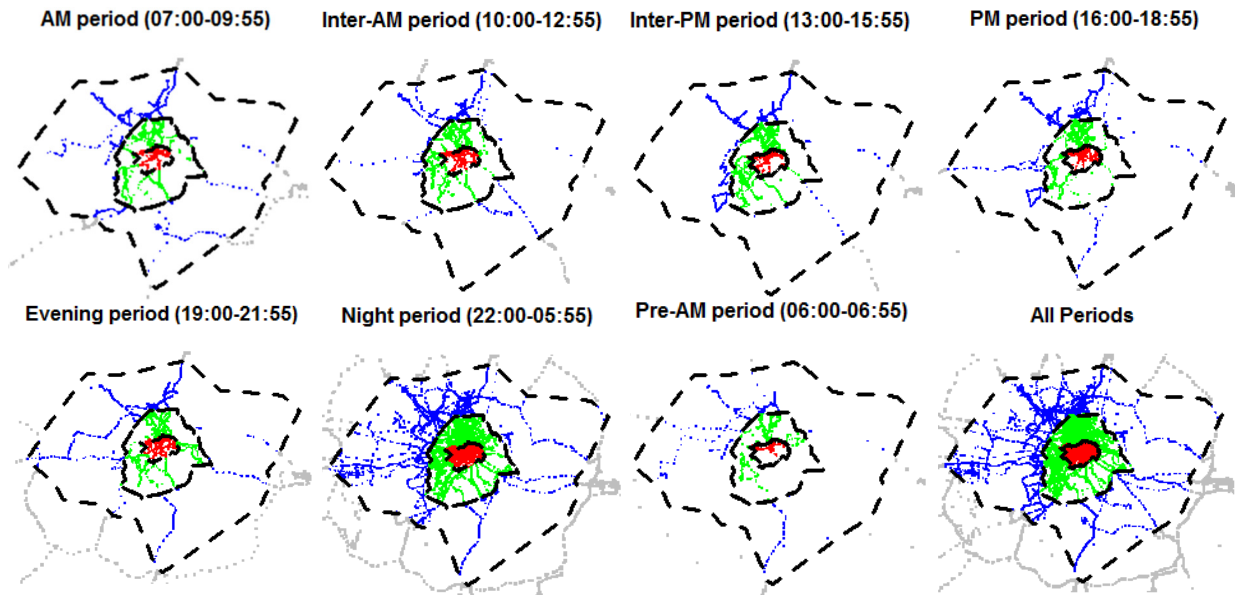


Figure 4.17 Validation Dataset Distribution for TfL Daily Time Periods in London

Figure 4.18, on the other hand, focuses on movements Central London throughout the day. In general, the participant activity looks concentrated in northern parts of Central London. Commute patterns seem to emerge however between across the rest of the zone. The data as a whole seems to reflect an expected pattern given the typical movement within London at different times of the day for a dataset with a significant number of participants that work/study near UCL area.

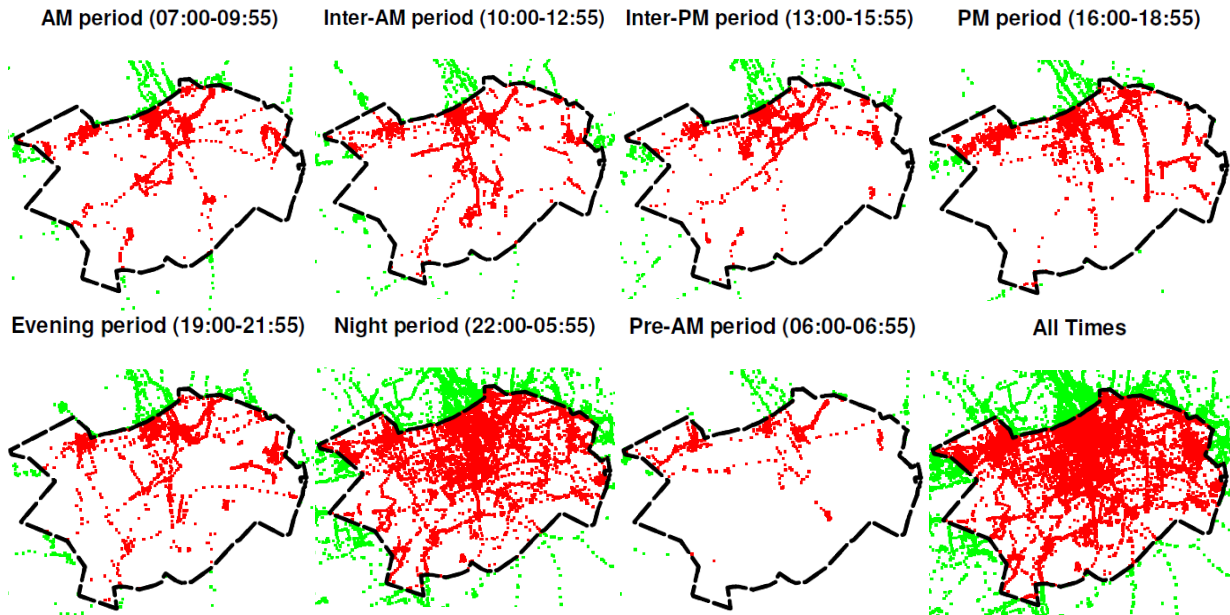


Figure 4.18 Validation Dataset Distribution for TfL Daily Time Periods in Central London

Figures Figure 4.19, Figure 4.20 and Figure 4.21 present heatmaps of the dataset distribution within Central, Inner and Outer London respectively. As demonstrated in the previous subsection, Central London appears to contain hubs where movements move to and from such as Euston Square, Oxford Street, Paddington Station and King's Cross Station. The temporal aspect highlighted in the heatmaps appears to stress on the occurrence of the activity around different hub areas. For example, the AM and PM periods seem to have most activity happening near major train stations such as Paddington and King's Cross. On the other hand, Inter-Am and Inter-PM periods seem to contain activity nearer to UCL and Oxford Street area. The night period however shows that movement occurs in different areas across Central London with more stress on hub areas and night-live zones such as Piccadilly Circus and Soho.

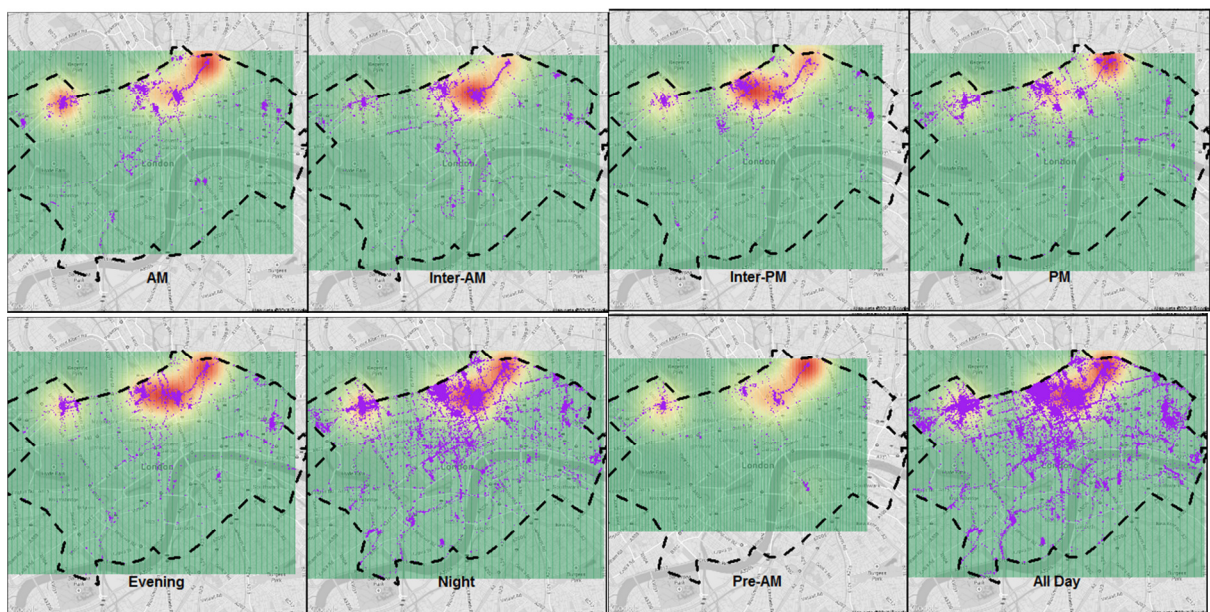


Figure 4.19 Validation Dataset Heatmaps for different Times of Day in Central London

Inner and Outer London also exhibit similar patterns where main hubs occur such as around Hammersmith Station for example. The dataset however reflects more concentration of movements in North London than anywhere else as could be noted from Outer London heatmaps. This is highlighted as a limitation of the collected validation dataset and must be considered when using results from work done in this thesis.

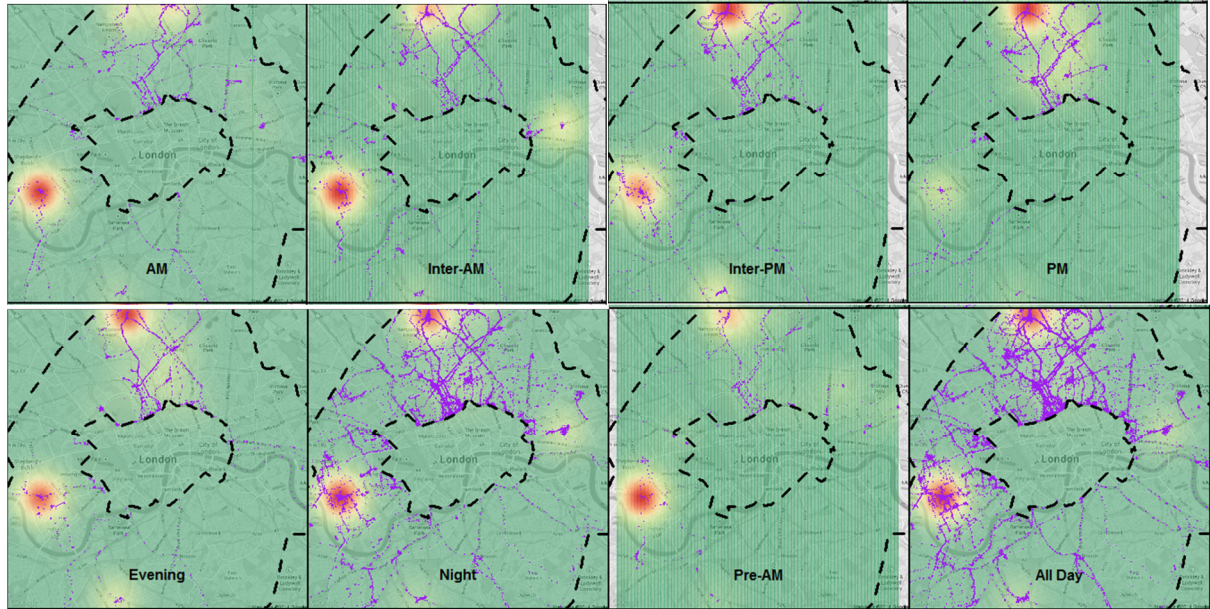


Figure 4.20 Validation Dataset Heatmaps for different Times of Day in Inner London

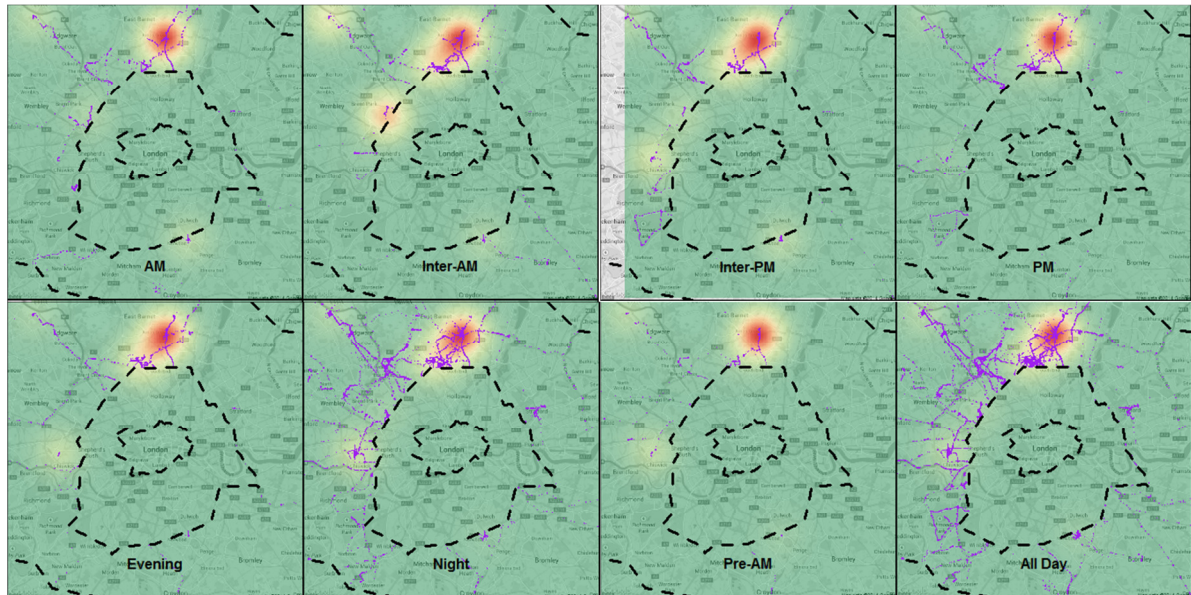


Figure 4.21 Validation Dataset Heatmaps for different Times of Day in Outer London

4.4.2.3 Spatio-Modal Distribution

This subsection discusses the distribution of the validation dataset based on the transportation modes used by the participants. Table 4.7 shows that stationary and walk modes encompass most of the dataset activity as might be expected. The stationary mode is

shown to constitute the majority of the states/activities of data of this study's participants (87%). This mostly reflects the dominance of periods of rest at the workplace, home or social hubs with respect to other daily activities. The walk mode (8%) is noted to be more concentrated in the Central and Inner London reflecting the higher interactivity level of road links with people using the streets in these areas.

| Mode | Central | Inner | Outer | Outside | Total |
|------------|---------|--------|--------|---------|--------|
| bus | 0.20% | 0.30% | 0.31% | 0.05% | 0.86% |
| car | 0.23% | 0.62% | 0.42% | 0.91% | 2.18% |
| cycle | 0.23% | 0.95% | 0.32% | 0.01% | 1.51% |
| stationary | 24.44% | 28.98% | 24.19% | 9.04% | 86.65% |
| train | 0.04% | 0.09% | 0.11% | 0.48% | 0.72% |
| tube | 0.06% | 0.12% | 0.16% | 0.00% | 0.34% |
| walks | 3.01% | 2.87% | 1.30% | 0.59% | 7.77% |

Table 4.7 Spatio-Modal Data Distribution in London (Colour Intensity-Coded according to Distribution Share)

Stationary and walk modes are usually regarded as ancillary modes, i.e. transition states between one mode and the following. Therefore, Table 4.8 disregards them from the analysis, reflecting the modal split based on non-ancillary modes across the spatial domain. The car mode constitutes around one third of non-ancillary modes (39%) probably due to the length and time of driving trips. The car mode also seems to be concentrated outside of London and more in Inner London than in other parts of London. Within London, Cycling mode (27%) seems to be concentrated in Inner London as well. This might be attributed to the sheer amount of daily travel occurring between parts of Central London and Inner/Outer London. Other modes such as train and tube appear to be concentrated where the network extends and where it is over-ground, as expected.

| Mode | Central | Inner | Outer | Outside | Total |
|-------|---------|--------|-------|---------|--------|
| bus | 3.58% | 5.31% | 5.48% | 0.88% | 15.25% |
| car | 4.07% | 11.00% | 7.45% | 16.24% | 38.76% |
| cycle | 4.04% | 16.93% | 5.68% | 0.17% | 26.82% |
| train | 0.72% | 1.62% | 1.92% | 8.67% | 12.93% |
| tube | 1.13% | 2.15% | 2.89% | 0.07% | 6.24% |

Table 4.8 Spatio-Temporal Data Distribution in London excluding Stops & Walk Modes (Colour Intensity-Coded according to Distribution Share)

Figures Figure 4.22 and Figure 4.23 help visualise the above distribution on a map of London's zones and Central London respectively representing different types of modes of transport. As expected, stops data is patchy and indicates the instances of inactivity which usually refers to places of interest to participants of this dataset. Walk data also refers to movements towards and from these hubs indicated by stops.

As for non-ancillary modes, some seem to be evenly distributed across different parts of the three congestion zones of London such as car and train. Bus and cycle modes on the other hand appear to be evenly-distributed only in Central London and clustered in different parts of outer London zones. Clusters in outer London zones occur as a result of the daily commute

routes of specific participants into Central London from their homes. This indicates a slight bias in the spatial representation of certain areas (especially Outer London) of bus and cycle modes.

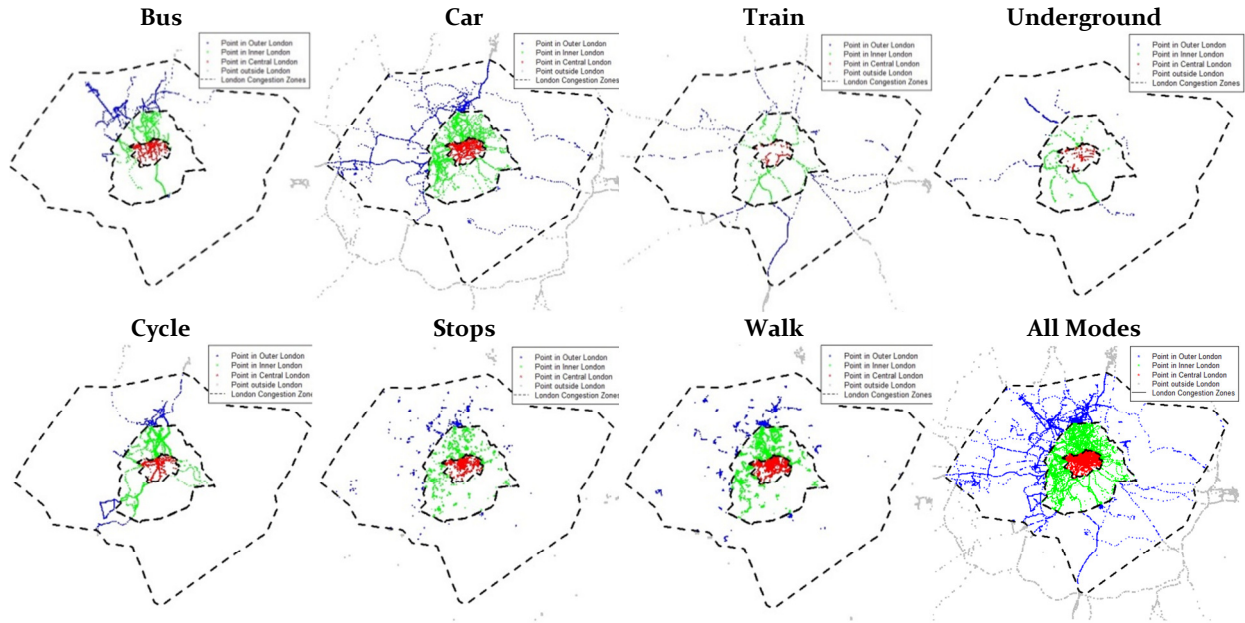


Figure 4.22 Validation Dataset Distribution for different Modes in London

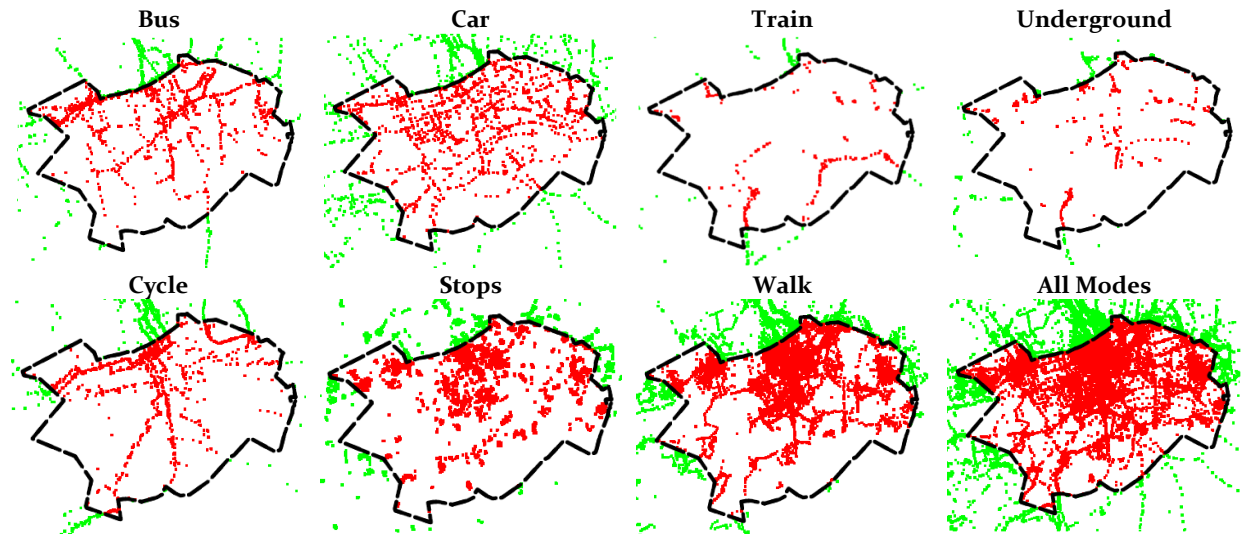


Figure 4.23 Validation Dataset Distribution for different Modes in Central London

Figures Figure 4.24, Figure 4.25 and Figure 4.26 help visualise the data above in the form of heatmaps for different modes of transport grouped by different zones of London. Train and tube data appears to be clustered around their networks, more specifically at parts of the network which are over-ground, which is constrained by the network topology and spatial extent. Car movement has been noted in the previous figures to be spread around the city of London; however in the heatmaps below it appears to be spread evenly only across Central London yet more clustered in areas of the north and west of Inner and Outer London. In general, bus and cycle modes appear to be clustered along certain routes which are mostly towards the north of all zones of London, while ancillary modes (walk and stops) appear to be densely clustered around specific transport hubs and participant-specific places. These

clusters are important to note as the extent of limitations of this study's performance measurement.

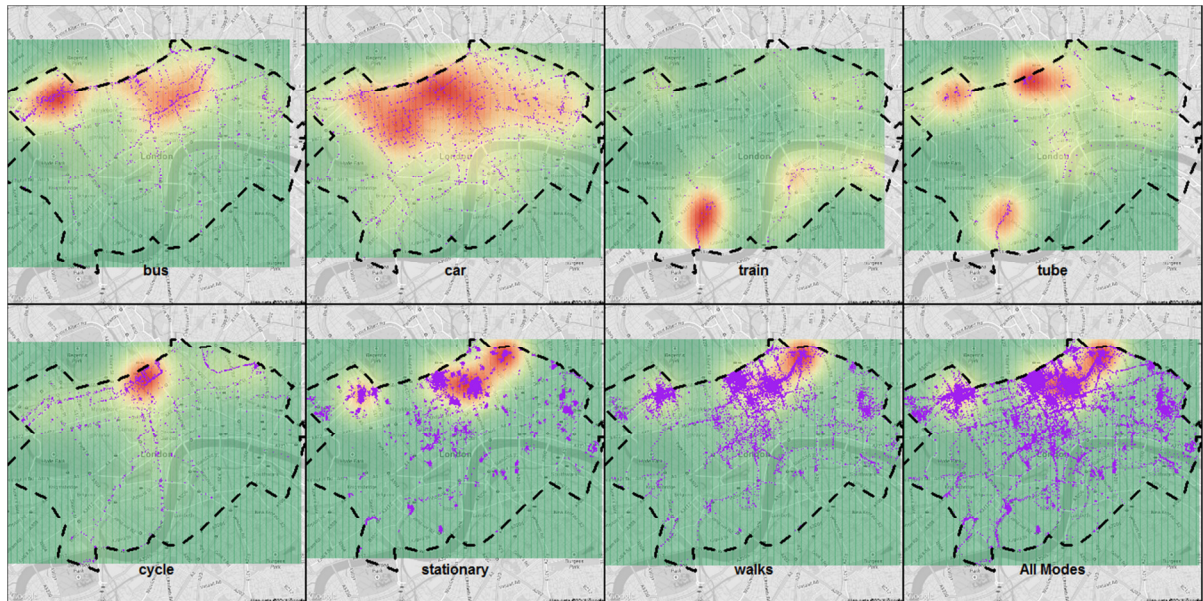


Figure 4.24 Validation Dataset Heatmaps of different Modes in Central London

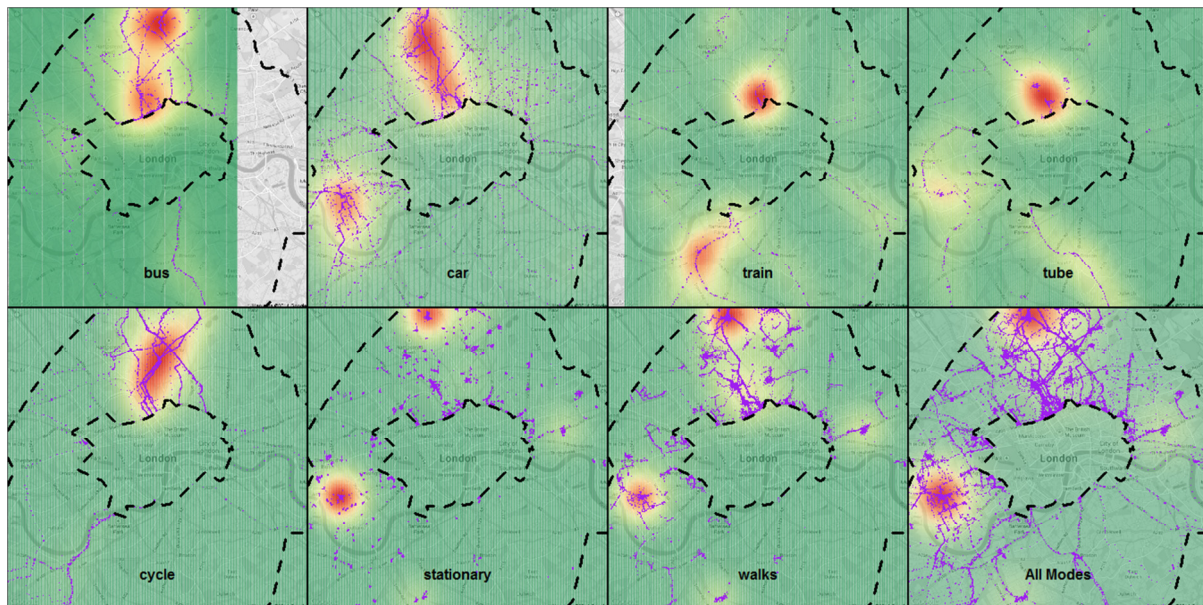


Figure 4.25 Validation Dataset Heatmaps of different Modes in Inner London

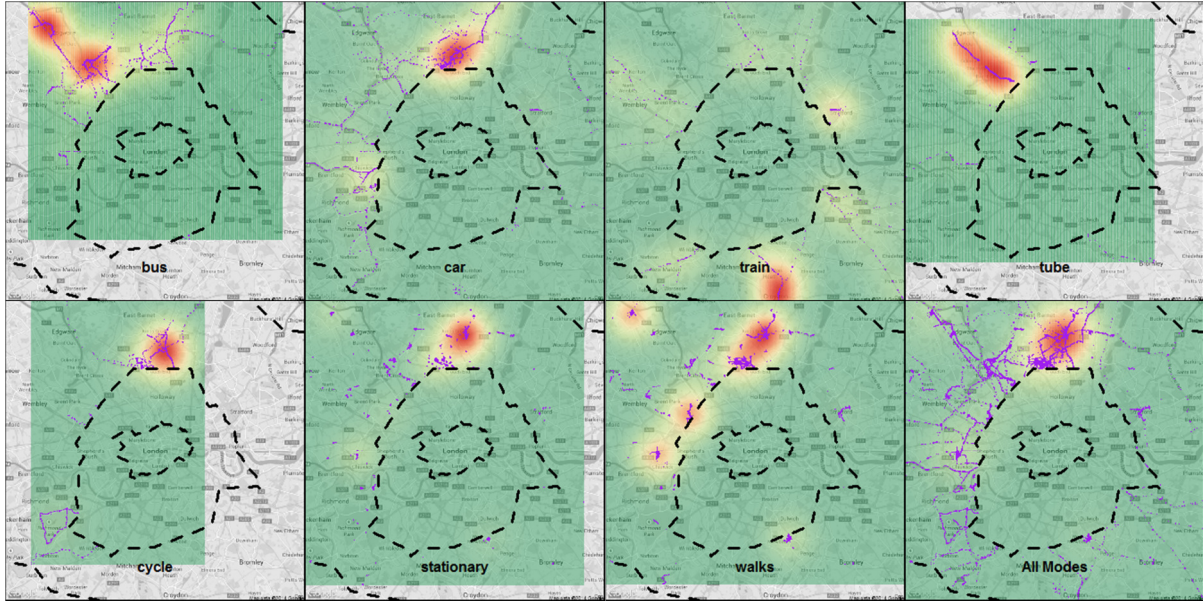


Figure 4.26 Validation Dataset Heatmaps of different Modes in Outer London

4.4.2.4 Modal-Temporal Distribution

Table 4.9 below shows a tabular distribution of modes relative to different TfL periods of the day. As might be expected, the stationary mode (stops) constitutes more than 85% of the dataset with one third of it occurring at night. On the other hand, most of the activity (non-stationary) seems to be occurring in the middle part of the day (Inter-AM, Inter-PM and PM).

| Mode | bus | car | cycle | stationary | train | tube | walks | Total |
|----------|-------|-------|-------|------------|-------|-------|-------|---------|
| AM | 0.15% | 0.39% | 0.63% | 10.67% | 0.10% | 0.09% | 1.14% | 13.17% |
| Inter-AM | 0.14% | 0.42% | 0.06% | 9.47% | 0.10% | 0.05% | 1.20% | 11.44% |
| Inter-PM | 0.16% | 0.35% | 0.09% | 9.40% | 0.11% | 0.03% | 1.19% | 11.33% |
| PM | 0.22% | 0.45% | 0.55% | 9.25% | 0.20% | 0.11% | 1.24% | 12.02% |
| Evening | 0.10% | 0.31% | 0.11% | 11.00% | 0.14% | 0.04% | 0.81% | 12.51% |
| Night | 0.07% | 0.17% | 0.03% | 32.82% | 0.05% | 0.03% | 1.95% | 35.12% |
| Pre-AM | 0.02% | 0.07% | 0.03% | 4.04% | 0.03% | 0.00% | 0.23% | 4.42% |
| Total | 0.86% | 2.16% | 1.50% | 86.65% | 0.73% | 0.35% | 7.76% | 100.00% |

Table 4.9 Modal-Temporal Data Distribution in London (Colour Intensity-Coded according to Distribution Share)

Table 4.10 therefore dismisses the stops data from the dataset to build a better understanding of the modal activity in the dataset. It could be noted from the table that most of the activity occurs in the AM and PM periods as might be expected due to daily home/work commute. The difference between these two peaks and the period between them (Inter-Am and Inter-PM) seem to be minimal for bus and car modes yet maximal for cycle mode. This might be attributed to using buses and cars to commute to work-related locations during work hours yet not using bicycles for such purposes. Another possible cause is that participants that cycle to work belong to a group of a work type of a static nature (office or similar).

| Periods | bus | car | cycle | train | tube | Total |
|----------|--------|--------|--------|--------|-------|---------|
| AM | 2.68% | 7.05% | 11.19% | 1.73% | 1.63% | 24.28% |
| Inter-AM | 2.50% | 7.47% | 1.11% | 1.86% | 0.85% | 13.79% |
| Inter-PM | 2.80% | 6.33% | 1.62% | 1.89% | 0.48% | 13.12% |
| PM | 3.92% | 7.99% | 9.86% | 3.52% | 2.00% | 27.29% |
| Evening | 1.82% | 5.52% | 1.94% | 2.44% | 0.79% | 12.51% |
| Night | 1.24% | 3.12% | 0.60% | 0.87% | 0.48% | 6.31% |
| Pre-AM | 0.29% | 1.29% | 0.49% | 0.62% | 0.02% | 2.71% |
| Total | 15.25% | 38.77% | 26.81% | 12.93% | 6.25% | 100.00% |

Table 4.10 Modal-Temporal Data Distribution in London Excluding Stops (Colour Intensity-Coded according to Distribution Share)

4.4.3 Limitations of Sample Properties

The extent and coverage of any sample collected for a study carried out in an urban setting such as London will always depend on places that are significant to participants of the study. Choosing the sample for this study has been carried out using stratified sampling where each subpopulation (stratum) held homogeneous sizes over a period of two weeks to account for seasonal variation. The stress was however on participants who worked/studied for long periods in a day yet keeping the strata collectively exhaustive where no population element is excluded. More detail regarding the sample composition and analysis is demonstrated in chapter 9.

The sample however was not chosen based on any spatial coverage of the data. Since the study has spread stemming from across the UCL network, most of the participants worked/studied in Central London or close to it. Upon designing the experiment, it was thought that such a sample would reflect the centralisation of big cities such as London, and the change in congestion levels over different periods of the day. In this section, we have tried to understand the spatial, modal and temporal extents of the data within London in an attempt to appreciate the coverage limitations of the dataset in these three domains. Coverage limitations are an inevitable by-product of any experiment, however clearly highlighting these limitations puts the study into its full perspective when adopting its results in any further work.

Among the limitations that this section has highlighted is that more concentration of movements in North and West London than anywhere else for the TfL congestion zone of Outer London. There is also a slight bias in the spatial representation of certain areas (especially Outer London) of bus and cycle modes. Another issue is the concentration of a portion of the data around the UCL area due to the origin of the network where participants were selected. Other issues such as the concentration of certain modes near transportation hubs and cycling occurrence in peak commute hours but not midday in Central London only reflect the nature of trend movements in complex networks such as London's.

4.5 Summary of data Collection Issues Addressed in this Chapter

This chapter has discussed the issues that affect the data collection phase of GPS-based travel surveys aiming at detecting the mode of transport. We achieve that by standardising the data collection phase through addressing four issues, namely; selecting an appropriate positional sensor technology and device to collect the data, selecting a well representative study area, understanding the sample spatio-temporal distribution and sample duration, and identifying the most representative collection rate for the selected GPS devices.

This chapter started by discussing the different types of positioning techniques along with their advantages and disadvantages. In the context of the device types suitable for detecting the mode of transport, we select two devices for pilot and validation data collection that possess the desirable properties for this kind of research. The chosen devices are based on GPS technology with a low battery consumption, high accuracy, non-real time, wide coverage and small in size. The two chosen devices are developed by u-blox and GTrek. The two devices are portable and small in size (5X3X1 cm). They both have a large battery capacity that lasts for almost a couple of months and a week respectively; to be able to track users in all travel modes in an unbiased environment. They are also passive (non-interactive hardware), for battery life purposes and to keep the user's attention away from it to insure an unbiased behaviour. The processing mode is of an offline mode (post-processing) to take off the burden of data entry every step along the user's journey. One of the devices (u-blox) is also equipped with an accelerometer that activates the data collection mode only when motion is detected. This helps increasing the battery life significantly. This device was developed by u-blox which contributes in funding and industrial supervision of this very research.

Another issue discussed in this chapter is the selection of an appropriate study area for conducting this study. As a result, London is selected as the testing study area. The data is collected for participants that are based or work in London, meaning that they conduct daily trips within London. London is selected due to its complexity and the diversity of its transportation networks.

The third issue is the sample collection rate where this chapter carried out an experiment aimed at identifying the most appropriate epoch rate of collection for GPS devices for travel behaviour studies purposes. The experiment concludes that the most appropriate epoch rate for route and origin/destination detection of pedestrians is probably somewhere between 30 and 60 seconds according to the datasets used in this study. Using this recommendation could arguably increase the accuracy of detecting the mode of transport from GPS data while minimising the battery consumption and computational cost. Further work could include running more experiments to conduct a similar study for other modes.

The final issue is the sample distribution across space, activities (modes) and time and the duration of GPS data to collect to represent movement in London adequately. Understanding the data distribution along these domains highlights the limitations that need to be accounted for when further implementing and analysing the results from this study. One limitation of the dataset is the occurrence of a concentration of movements in North and West London in the area of Outer London to some extent. Outer London also holds a slight spatial representation of bus and cycle modes. On the other hand, the occurrence data clusters

around transportation hubs (train stations, underground stations, bus stops) and the exhibition of certain modes at certain areas during certain periods of the day only reflects the pattern of life and typical general movement trends that a city would have.

As a result, the two datasets collected for the purpose of validating the classification framework developed in this thesis are described as follows:

1. ***The pilot data*** – 21 people over 2 weeks using u-blox GPS devices (u-blox, 2009) at a collection rate of 60 seconds; where the mode of transport is labelled by the participants segment-by-segment,
2. ***The validation dataset*** – 84 people over 2 months using GTrek GPS devices (GTrek, 2012) at a collection rate of 30 seconds; where modal information and ownership of a bike, access to Barclays Bikes, driving license and ownership of a car is provided by the user.

As mentioned in chapter 2, the data collection issues are one out of five main issues this thesis addresses in the context of detecting the mode of transport from GPS data. The following four chapters describe the remaining issues addressing the topic.

Chapter 5

Data Pre-Processing

5 DATA PRE-PROCESSING^{8 9}

The accuracy and standardisation of metadata associated with a GPS track defining the type and level of information affects the process of validation of the inference framework developed in this thesis greatly. The quality of this information also attributes hugely to the whole validation process, as can be summarised by the phrase “garbage in, garbage out” being applicable to any data mining and machine learning project (Pyle, 1999). Therefore, defining a positional data pre-processing session depends entirely on the level of detail required for the study and the quality by which it is created. In this chapter, we address the limitations in the data pre-processing phase for GPS-based travel surveys aiming to detect the mode of transport. We mainly discuss four issues highlighted in chapter 2 in the data pre-processing phase (as shown previously in Figure 2.3). These issues include data cleansing, choice of modes, data generalisation and data labelling. This chapter is broken down into several sections each describing one or more of these issues.

We start by addressing the GPS data accuracy limitations due to low positional accuracy and signal loss problems in section 5.1. The section also discusses the modes of transport to be studied within the scope of this research.

Section 5.2 then defines the level of detail and granularity of the collected GPS tracks’ information (or metadata) that will be assessed in this research. This section also attempts to standardise the process of attaching metadata by the participants when labelling their tracks, which standardises the metadata (mode of transport) using which the classification method will be tested and validated based on.

Finally, section 5.3 discusses reasons why traditional travel diaries return low response rates or inaccurate information from the side of participants taking part in these surveys. The quality and quantity of the feedback information provided by the participants affects the quality of the method and validation process of the classification framework developed in this research (and any other research). Therefore, the section then introduces an online web-based interface through which participants can upload, view, and edit their tracks overcoming all the travel diary limitations that provoke low and inaccurate response rates.

5.1 Managing GPS Errors and Choice of Modes

As discussed in chapter 3, we explain the guidelines and practices previous research used to deal with and benefiting from GPS limitations by filtering and processing GPS data. We demonstrated that there are two cases of implications of GPS technology limitations, namely; low positional accuracy and signal loss. This section describes how we attempt to deal with

⁸ Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T, Tsapakis, I & Haworth, J, 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems*, Special Issue: *Advances in Geocomputation*, Volume 36, Issue 6, pp. 526–537.

⁹ Part of this chapter was presented in the following publication: Bolbol, A, Cheng, T, and Paracha, A, 2010. GEOTRAVELDIARY: Towards Online Automatic Travel Behaviour Detection. *WebMGS: 1st International Workshop on Pervasive Web Mapping, Geoprocessing and Services*, Politecnico di Milano, Como, Italy, August 26–27, 2010.

these GPS limitations in a correctional and a beneficial manner. This section also describes our strategy on selecting the modes of transport to include and account for in this study.

5.1.1 Low Positional Accuracy

Low GPS data accuracy is usually due to systematic errors caused by one or many situations such as having too few satellites in view, having the satellites aligned in a bad geometry (High HDOP), multi-path signal reflection, Ionosphere effects, signal blocking, or by clock or receiver issues (Jun, et al., 2006). As described in chapter 2, several research attempts have been carried out to address these GPS inaccuracy errors such as *data filtering*, *data smoothing* and *identifying and excluding stops*. The following subsections describe our attempt to address these issues in this research in relation to strategies implemented in previous research.

In this work, we follow some of previous research practices when dealing with low positional accuracy problems such as wandering errors & urban canyons by **filtering** data by dismissing fixes with less than 4 satellites, HDOP more than a value of 5, unrealistic altitude value, speeds above 50 m/s or accelerations above 10 m/s². However, we do not apply any strict GPS data smoothing techniques for eliminating urban canyon effects in order not to change the raw nature of the data and the speed and acceleration values calculated from it. The importance of preserving the positional rawness of the data stems from the fact that our methodology carries out processes such as classification, clustering and network matching based on this data, and any intended changes to the data might significantly bias the results.

5.1.2 Signal Loss

As also mentioned in chapter 2, signal loss is categorised into partial and total signal blockage. *Partial blockage* occurs when less than 4 satellites are captured in the fix and that will be due to an urban canyon or usage of public transport. Conversely, a *total blockage* occurs when no satellites are captured at all and happens upon entering indoors, underground or into a tunnel.

Previous research has attempted to overcome **partial blockage** by integrating the GPS device with other devices such as gyros, Wi-Fi, or GPRS technologies. The disadvantage of these hybrid solutions is that such equipment is usually expensive, less portable, or consumes high battery power. Therefore, we only use pure GPS technology for collecting positional data. We use filtering techniques described in the beginning of this section to filter out partial blockage using HDOP, satellite, altitude and speed information of every GPS fix.

On the other hand, previous research has treated **total blockage** by using transport networks to find missing segments of a GPS track (Stopher, et al., 2005). Total blockage could be also beneficial for identifying the certain modes such as underground travel. In this research, work attempting to identify underground travel is described in detail in chapter 8 as part of describing attempts to match GPS data to corresponding transport networks leading to the inference of the used mode of transport. Total blockage has also been used in previous research to find trip ends; leading to the segmentation of tracks into trips (Stopher, et al., 2008a). We also adopt this strategy to identify individual trips within a GPS track. This is

described in detail in chapter 7 discussing our attempt of carrying out the whole process of track segmentation.

5.1.3 Selected Modes of Transport

The validating and testing of algorithms that aim to infer the mode of transport are only as accurate as the information provided with the data. This information includes the types of mode of transport that are to be investigated, the generalisation level of the trip information, and the strategy used to report the mode of transport by the participants. As discussed in the rest of this section, some decisions made regarding this information hold many limitations. Among these limitations is considering a limited number of modes of transport for the developed method. This makes these methods less robust to identify ignored modes. Therefore in this study, we consider all 6 modes that are used on different transport networks in London. These are: (1) bus, (2) cycle, (3) car, (4) train, (5) underground and (6) walk. One can also argue that stops can be considered as a mode on its own as “stationary” for example, however in this work, it is only considered as an independent phenomenon that often occurs within any GPS track. The following section describes what we define as a stop.

5.2 Trip information Generalisation

Another limitation that some studies possess is generalising the modes used in a trip to only the most dominant mode within this trip (Stopher, et al., 2008a; Manzonni, et al., 2010). This decreases the accuracy of the learning and validation process as a result of having a mode being denoted by mixed modes. An example of that is having a trip with a train and bus stages, and a decision on counting the trip as once stage of train since it covers most the travelled distance for the trip while ignoring the bus and walk stages. Hence, standardising the components of a trip and identifying the decisions that will be made according to different movement situations will lead to overcoming such limitation. Therefore, the degree of generalisation of the metadata of a given trip has to be well-defined and standardised. Moreover, decisions on what mode of transport to assign while attaching the appropriate metadata to a GPS track in different defined scenarios also need to be standardised.

Therefore in this section, we describe how we define the metadata associated with each GPS fix. We also define different segments in the hierarchy of GPS track. We also explain why we establish the definitions in the way we do. We first set some definitions used in this research regarding the different elements of a GPS track. A hierarchy has been developed and set for this kind of definition to clarify and distinguish different parts of a trail. The process of segmentation and labelling is then discussed to set the guidelines for user-generated track labelling. As a result, a set of rules is developed to regulate this process to eliminate confusion. We describe these guidelines, along with labelling decisions in specific problematic scenarios.

5.2.1 Elements of a Track

A GPS track consists of many different parts (or elements) that are defined based on the main purpose of the application the track is needed for. In this research, fine details of a GPS track is of great importance since the study aims to understand travel behaviour and infer the mode of transport as a result. This means that any broad generalisation or extra unnecessary detail added to the track will influence the accuracy of perceiving behavioural elements from the

GPS track. Consequently, we define a hierarchical structure of the elements of a given GPS track. The structure is shown in Figure 5.1 showing the *User* in the highest level, since the study consists of more than one user. Every user uploads/produces more than one *Track* where each which might contain more than one *Trip*. From the *Trip* level downwards, all the elements are considered to be the main elements that influence a mode of transport inference model.

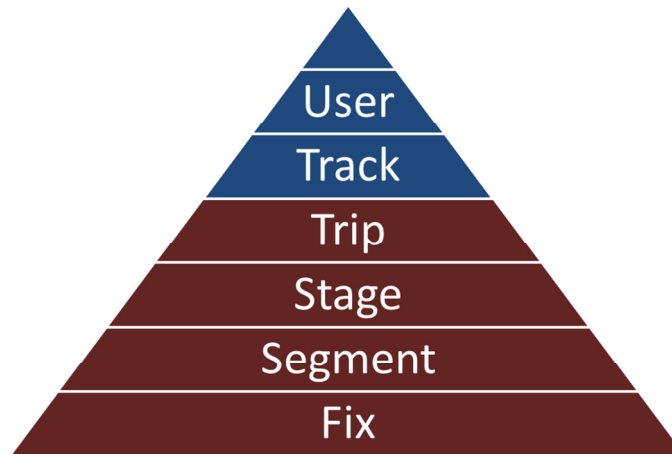


Figure 5.1 GPS Hierarchical Data Structure in this Research

As mentioned briefly in chapter 2, we standardise some definitions to be used for the description of different elements of the *trip* in travel survey studies. As demonstrated in Figure 5.2, the route between any two consecutive GPS points is called a *segment*. Trips also consist of a number of *stages* (a group of segments). A new stage is defined when there is a change from one mode of transport to another, or where there is a change in vehicle of the same mode (Anderson, et al., 2009). A point separating two stages is called a mode change point (also called mode switch). The first and last points of a trip are called trip ends or origin and destination of a trip, yet they will not be the focus of this research. Moreover, any cluster of GPS points within a stage is considered to be a stop.

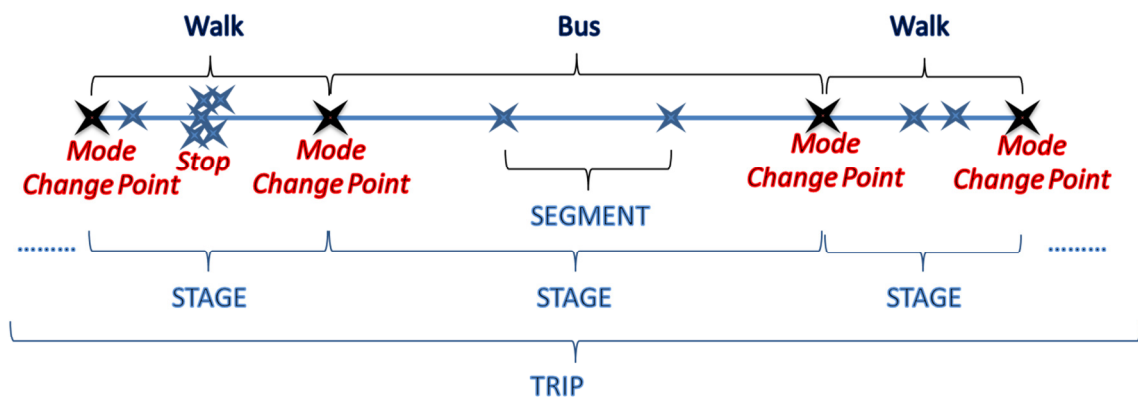


Figure 5.2 GPS Trip Elements

Figure 5.3 presents an example of a trip having an origin (point 1) and a destination (point 9). The trip consists of two stages one of which consists of 4 segments (A, B, C and D) bound by 5

points (1 to 5), where the mode of transport is cycle. The other stage consists of another 4 segments (E to H) bound by 5 points (5 to 9) where the mode is walk. As might be noted; point 5 exists in both the cycling and the walking stages; and hence is called a Transition Point (Zheng, et al., 2008) or a Mode Change Point as mentioned above.

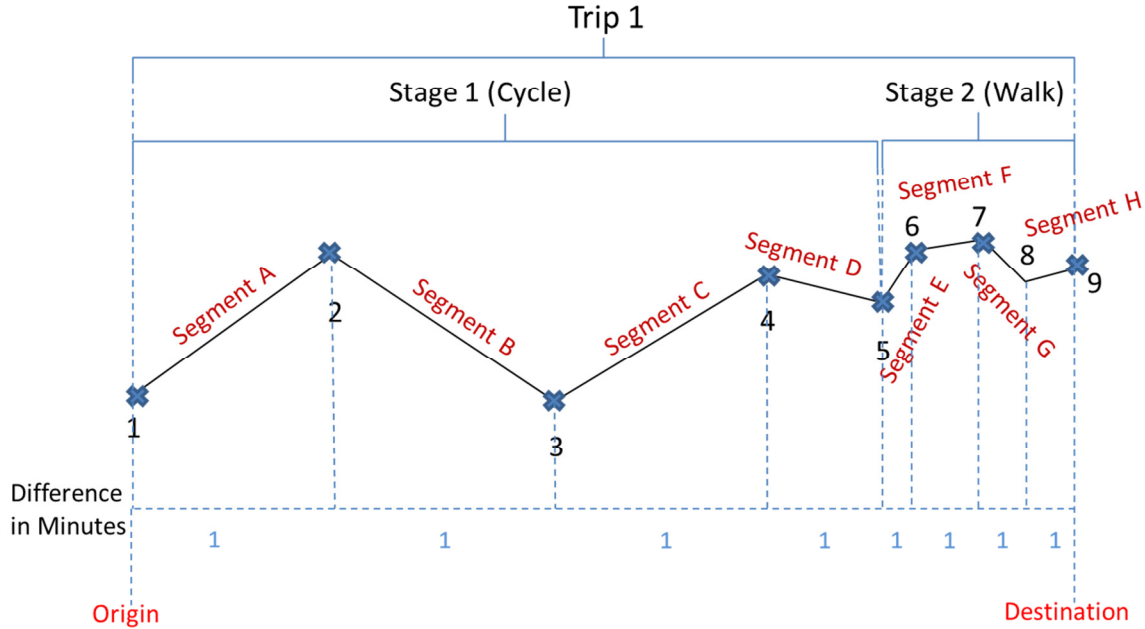


Figure 5.3 A Trip Description

5.2.2 Segmentation and Labelling

The segmentation process is a very essential part of the data pre-processing stage. Defining the rules for segmenting a track into its constituents is an essential step in metadata production in order to ensure that different tracks are labelled based on the same rules, and hence, eliminating discrepancy in the learning and the validation phases. Therefore, this section describes the rules we assign to unify the track's labelling decisions at confusing situations in order to set guidelines for participants to label their tracks according to unified standards.

5.2.2.1 Segmentation Rules

A stage, which consists of a single mode, could be identified in various different ways. Therefore, we need to unify and set the definition of what is to be considered a stage and where exactly it starts and ends. The level of detail of defining segmentation rules during the labelling phase (attaching metadata) affects the outcome classification obtained, and hence, affecting the validation process and the achieved accuracy. Thus, a couple of preliminary rules were decided upon for defining a new stage for the purpose of this research. A new stage is formed whenever:

- the mode of transport changes,
- a stop is committed for one of the stop purposes
- or a destination is reached

As mentioned in chapter 2, a long standing argument would be what to count as a stop (Stopher, et al., 2005). Standardising the rules by which participants will tag stops in their track, we will be able to accurately assess the efficiency and accuracy of identifying these stops.

5.2.2.2 Rules for Significant Stops

Therefore, we have created several rules for significant stops (or trip-ends) based on TfL (2009b) and R.A. Malatest & Associates Ltd (2012). Table 5.2 summarises these occurrences showing different possible trip purposes and significant stops that are expected in such trips. For example, if a trip's purpose is sports/leisure then the significant stop or trip end is defined as the first GPS fix obtained at the pitch/playground/gym/swimming pool/etc.

| Purpose | Significant Stops |
|-------------------|---|
| Return Home | Home |
| Work | Workplace |
| Education | Educational Institution |
| Shopping | Shop |
| Personal Business | Destination |
| Leisure | Pitch/Playground/Gym/Swimming Pool/etc. |
| Recreational | (If it's a park walk or cycling for example then the end is the same as the origin) |
| Social | Destination |
| Dining | Restaurant/Takeaway/etc. |
| Pick up | Pick up Location |
| Drop off | Drop off Location |

Table 5.1 Rules for Stop Locations Based on Trip Purpose

5.2.2.3 Rules for Stops

Another problem that participants would face while labelling their tracks is what classifies as a stop, since other modes usually include occasional stops along the track. Many questions arise such as (Do we classify walking then waiting for the bus as a stop?) Or (is stopping part of the walk stage?). As previously mentioned, these decisions entirely depend on exactly how much detail is needed for the classification. In order to overcome these limitations, there is a need to produce standardised guidelines for taking decisions on stops when tagging a track. For that reason, we have also created rules to define stops in the case of the most common scenarios that usually occur within a typical urban trip. These rules are described in Table 5.2 illustrating situations such as at bus stops, car parks, shopping, and picnics, etc. The table also describes in which case a decision would be made to include the stop as part of a stage with a certain mode of transport or a stop.

| Scenario | Mode of Transport | Stop |
|---------------------|---|--|
| Bus stop | If the wait is <2 fixes, the stop is considered as part of previous mode | Only if the wait is >1 fix (1 minute) |
| Outdoor Picnic | - | Always considered as a stop |
| Building | If the wait is <2 fixes, the stop is considered as part of previous mode | Only if not a stop and for less than the specified threshold |
| Bike chaining | If the wait is <2 fixes, the stop is considered as part of cycling | Only if the wait is >1 fix (1 minute) |
| Car parking | Car driving to destination & parking is all considered as car mode | - |
| Shopping | Motion while shopping or stopping and moving is considered walk mode (unless there was a long loss of signal (indoors) and unless its >1 fix) | Only if the stop is >1 fix (1 minute) |
| Appointment waiting | If the wait is <2 fixes, the stop is considered as part of previous mode | The wait is stationary (given its >1 fix) |
| Outdoor sports | It is all considered as walk | Only if the activity was of a static nature such as weight lifting, goal keeping, etc. |

Table 5.2 Decisions for Specific Labelling Scenarios

Table 5.2 bases all decisions on accounting for only transport movements, since the focus of this research is on understanding transport behaviour. Examples of the scenarios described in Table 5.2 that might be confusing when labelling GPS tracks with modes of transport is a “bus stop” situation as shown in Figure 5.4. A bus stop scenario occurs when participants walk to a bus stop to wait for the bus for some period of time and then change their current mode to “bus” as soon as they get on to the bus. Table 5.2 describes that a stop label is only considered when the waiting time at the bus stop is more than 1 fix long (meaning at least 1 minute). A similar scenario occurs when a participant cycles to a destination, ties his bike (taking some time to do that) then continuing his journey by walking or so forth. The decision in these two scenarios and similar scenarios mentioned in Table 5.2 is that the period spent waiting for the bus and chaining the bike are considered to be a stop. This ensures a higher accuracy of the labelling process and ensures that there would be an activity at that stage that will be inferred easily using this decision.

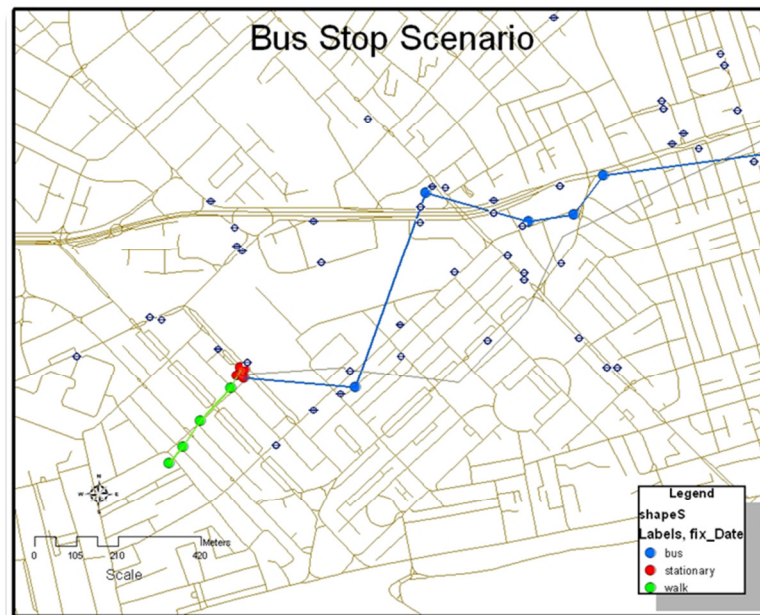


Figure 5.4 “Bus Stop” Scenario Example

Other examples from Table 5.2 include the “appointment waiting” and “sitting outdoors” scenarios. They both involve a participant moving towards a destination, stopping for a while then starting to move again. The decision is the same as the previous two examples, however, if these waits were the main destination of the trip. Figure 5.5 illustrates an example of these an appointment waiting scenario.

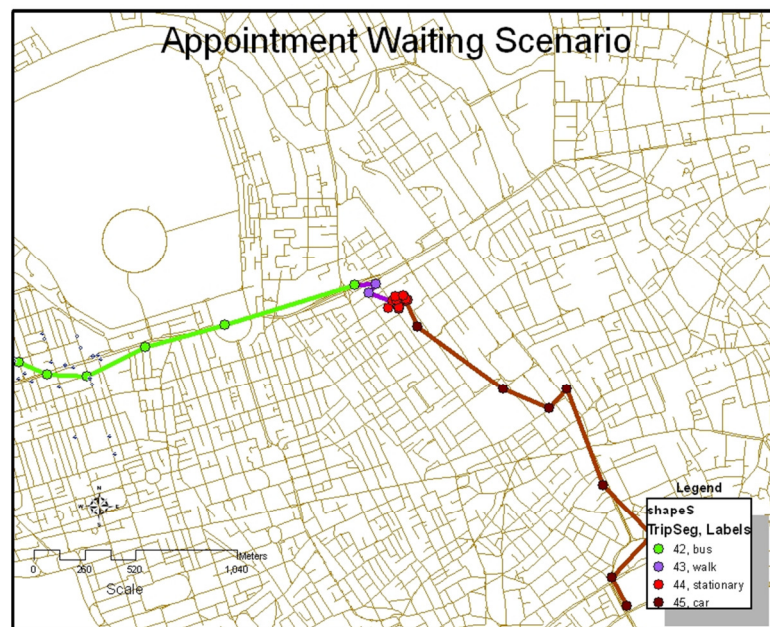


Figure 5.5 Example of an “Appointment Waiting” Scenario

On the other hand, a “car parking” scenario would all be labelled as one segment containing one more (including the parking activity). That is because the participant is still in the car driving whilst parking, so it is more considered to be a car mode than a stop. In this instance the car mode will include some stops such as at traffic lights. A similar situation occurs in a

“shopping” scenario where a participant moves between shops in an outdoor environment. A decision is made that if any of these stops exceeds 1 minute (having at least two fixes) then it is counted as a stop. Moreover, if the participant moves indoors for more than the specified threshold (e.g. 15 minutes) or if it is the main destination of the trip; then it would be counted as a stop. Conversely, an outdoor picnic scenario is usually meant to have the destination as where the picnic occurs, and will also be considered as a stop since it is an outdoor activity with GPS fixes. An example of an outside scenario is shown in Figure 5.6.

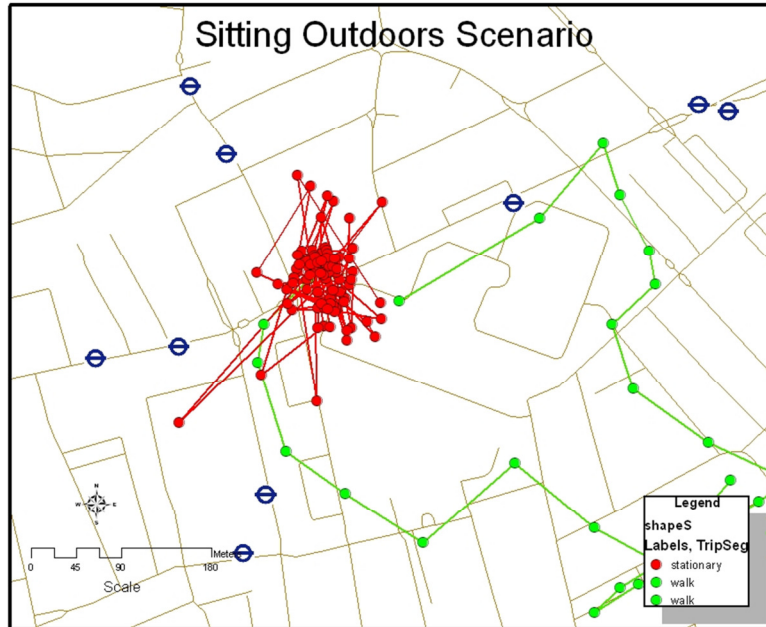


Figure 5.6 Example of a “Sitting Outdoors” Scenario

5.3 Data Labelling Strategy

The metadata, such as the mode of transport, used to validate the results of the inference method is usually added by the participants reporting the details of their GPS tracks. Some studies do not use participant-reporting to add this metadata to the GPS data track and instead depend on assumptions that may bear a large amount of inaccuracy. These studies do not collect participant feedback due to reasons such as the being burdensome and bearing privacy issues for the participant bears, as well as being slow and expensive for the data collector. Instead, some web-applications exist that enable the participants to view their tracks and edit them, yet still bearing several limitations such as not being able to add data metadata. Only one study enables the participants to fully check and edit their tracks. However, the study holds other limitations such as presenting unfamiliar basemaps, not using useful web-mapping products useful for better recall such as Google Streetview, and enabling the user to only use the data the data collector provides limiting the interaction of the participants with the data. For this reason, in this section we develop an online travel diary where participants can upload and label their tracks with metadata such as the mode of transport, as well as adjusting the position of some erroneous fixes their GPS devices collected.

Traditionally, travel diaries are considered to be a very important source of information that benefits major applications such as travel behaviour analysis. They have also been proven to be a burden on users to maintain and recall exact details, as well as being slow, expensive and time consuming. However, the rise of Geoweb 2.0, crowd sourcing and user-generated content is changing the way data is collected and shared. Nowadays, travellers could record their tracks and are able to upload them directly into the web to see them and even share them. The aim of this section is to develop an online user-generated travel diary system that records users' trails, which is also published in Bolbol et al (2010). The online server-side application allows the users to visualize their trails and manually edit, move or remove points from their tracks on a map-based interface. Such a map-based interface enhances the retrieval of travel information from the users' long-term memory. The interface also allows them to add such information as metadata to different parts of their trip, such as the purpose of the trip and the mode of transport. The application also overcomes travel diary disadvantages by being faster, cheaper and by providing users with incentives provoking higher response rates. This in turn will compile a large database of travel information within a large city (such as London in this case). Therefore, this will provide a better understanding of travel behaviour, and hence providing a platform for developing a realistic algorithm for learning travel metadata and hence deducing it.

5.3.1 Introduction into Non-Expert Online Travel Diaries

As mentioned in this research, location data is collected using GPS devices. Along with location data, metadata like the purpose of the trip and mode of transport is collected by participants using travel diaries during or after the journey. The GPS data and the meta data has been traditionally collected using travel diaries. The feedback has been proven to be a burden on participants to provide while filling travel diaries and many other participants do not recall exact details, as well as being slow, expensive and time consuming (discussed in subsection 5.3.2). Therefore, a cheaper, faster, and inviting alternative method is needed to overcome these disadvantages. There is also a need to understand how people record certain

events in their long-term memory and how they best recall them (discussed in subsection 5.3.3). Moreover, there is also a need to understand what intrigues people to participate in the travel information collection in order to obtain better response rates (discussed in subsection 5.3.4).

As the whole world is going online, many long processes that used to be carried out very difficultly (sometimes manually) need to be brought to a digital, more dynamic and interactive environment via the web. Among these is the collection of travel information. In this section, we suggest that overcoming all of these disadvantages could be by achieved by having an interactive, fast, and online system that takes advantage of all the new Geoweb 2.0 features. We also suggest using incentives that invoke user-generated content to be produced along with all the best methods of retrieving information from the human long-term memory without creating a sense of privacy invasion of users. The application GeoTravelDiary.com is developed and presented in this work and is discussed in subsection 5.3.5.

5.3.2 Travel Diaries

Travel information collection has proven to be a very hard process. Many examples have proven to be expensive, burdensome, dangerous, slow, difficult to plan, come back with inaccurate information or have low response rates. The following is a summary of the previous attempts to conduct different studies to collect travel data from either recruited or volunteered respondents. Along with that we shed some light on the various problems highlighted by different research.

5.3.2.1 Surveys, Diaries & GPS-Aided Diaries

One of the oldest yet still used methods is holding travel surveys either by telephone or face-to-face. This has proven to be a quite expensive, time consuming and sometimes dangerous to accomplish in some areas (Stopher & Metcalf, 1996). Another trend was to engage people by maintaining their own travel diaries through the telephone, paper and pen, or computer-based (Stopher & Greaves, 2007). It was reported however that travel diary surveys have very low response rates because they were a burdensome task and some users omit certain travel information due to fear of their privacy invasion (Wolf, et al., 2003). A new trend emerged as to use GPS devices in conjunction with traditional surveys (Stopher, 2008). Some research even analysed GPS data in order to minimize trip under-reporting through improved survey methods (Bricka & Bhat, 2006).

5.3.2.2 GPS with Prompted Recall Surveys

A further step was to base the diary on GPS devices and subject the respondents to undertake prompted recall surveys. Attempts have been made to process users' GPS data and provide maps to users of their tracks, mode of transport, and committed stops. They were then asked to visually verify and validate any identified errors on the map (Stopher, 2008). Prompted recall surveys were considered the best means to obtain good accurate results, where respondents had the opportunity to see their tracks on a street map background which helped them identify errors and misinterpretations. However, among the main problems with prompted recall surveys are that they still achieved very low response rates (Bricka & Bhat,

2006). And if responded to, the main problem was the respondent's accuracy in providing the inputs because they weren't very involved in the map making process which is a specialist process.

5.3.2.3 Real Time Diaries

With the advance in the new emerging technologies in the market place, other research attempted to use real time mobile technologies to collect location data and an input respondent-entered data together in the same process. This theoretically is the most practical and time saving method that a travel diary could undertake. On the other hand, many problems emerged due to real time usage, among these are the following.

1. **Privacy:** A sense of giving up one's privacy usually arises unless there is a significant amount of time between when the data is collected and when the respondents fill in the surveys (Stopher, 2008).
2. **User's Task Completion Time Gap:** The time lag between the instant the user starts on a different activity and the instant he/she records his/her status.
3. **User's Burden:** Users usually seem reluctant to record everything they do, because the real time activity takes some effort/time and occurs on several occasions along the day.
4. **Battery Requirement:** A common problem with real time GPS devices is having a short battery life.
5. **User Incentive:** Another very common problem that challenges lots of the previous research is the user's incentives. Some paid the users money (Stopher & Metcalf, 1996), while others tried to use special access panels that are set up with the purpose of conducting several different types of surveys on the same group of respondents by using incentives (Roorda & Miller, 2004). The rest mostly used volunteers that in many situations don't complete the full tasks (Stopher & Greaves, 2007).

In the application presented here, we try to overcome these shortcomings by tackling all the problems raised by the traditional travel diaries. The application takes the advantage of prompt recall surveys by providing an online map with their tracks requesting to provide feedback. It also is meant to be used in a desktop environment after the trips have been done in order to overcome the real time issues such as privacy, task completion time gap and the user burden. Users also have the incentive to participate by sharing and analysing their tracks (as discussed later in subsection 5.3.4.1). Long life GPS devices are also used to overcome the battery constraint (subsection 5.3.5.2).

5.3.3 Recalling Spatial Information

The way that the human brain collects, manages, organizes, stores and retrieves information is quite a fascinating yet puzzling process. It is a process that happens almost unconsciously in our everyday life. Understanding such a phenomenon and being able to maximize the brain's ability to perform these tasks; could significantly enhance the quality and speed of collecting accurate travel information feedback from users. In this section we discuss principles of

human cognitive psychology, and how to use it in favour of obtaining valuable feedback from users.

In a typical scenario of a travel diary, the different phases by which information is collected, stored and retrieved could be summarised in Figure 5.7. The figure illustrates two main phases; firstly, the data collection phase where the GPS data is collected and the user (consciously or unconsciously) uses his working memory to remember the most recent roads he/she took, then only the encoded information flows into the long-term memory. Secondly, the information is evoked to be retrieved when the user is interrogated about which routes and modes he took on his journey.

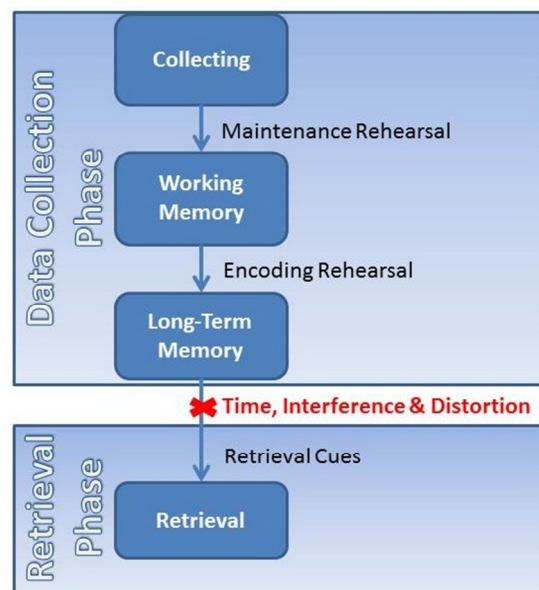


Figure 5.7 Information Flow in/out of the Human Brain

5.3.3.1 Data Collection: Information Encoding into the Long-Term Memory

The human brain attempts to rehearse information all the time. Psychologists distinguish between two types of rehearsal (Gray, 2001). Maintenance rehearsal is the process by which the brain holds information in the working memory for a period of time, and the encoding rehearsal is the process by which a person encodes this information into the long-term memory. Not all information that passes through the working memory is encoded into the long-term memory. Only information that is elaborated, organised or visualised is encoded. Elaboration means that we remember things that capture our interest and stimulate our thought. The information could also be encoded by organisation by chunking the amount of items to be rehearsed into fewer items. The more the experience with the similar information the easier is the correlation between the different items. This is a similar to the case with master chess players being able to memorise locations of all pieces on the board (De Groot, 1965). Information also could be encoded by visualisation of items in mental images.

When people travel, they experience the route they are taking in different ways. A good objective here is trying to pass as much information as possible into people's long-term memory. That is because according to Gray (2001), information in people's long-term memory may exist without them being aware of it, and using certain retrieval cues later, this information could be evoked. A natural advantage is that people usually elaborate on spatial navigation information because of the fact that they experience it in a real life context. People that take a similar route everyday (e.g. to work) tend to become experts of the area covered within the travel. This aids the objective of organising the route information and memorising connections between different streets. Also, much potential could be used to take advantage due to the fact that people tend to store mental images of their routes, by presenting users with photos of their different trip locations.

5.3.3.2 Causes of Human Information Loss

As illustrated in Figure 5.7, there are some causes of memory loss such as time, interference or distortion. Time is a main factor memory loss where as more time passes the more the memory loses the stored information. Ebbinghaus (1885/1913), however, argues that most of the forgetting happens in the first short period then the rate of loss becomes increasingly gradual. Therefore, the sooner the users log in their feedback the better, however, after a certain period (e.g. 2 days), the rate of forgetting stops increasing rapidly, and not a lot of information is lost onwards.

Another factor that contributes to forgetting is interference with one's memory. This relates to having to remember lots of other trips in the same area that users do. This is because the separately learned trips lose their distinctiveness and merge into one large pool (Gray, 2001). However, representing these trips on a map to the user, and having them segmented by time, loss of signal or by distance might overcome this problem.

The last factor causing memory loss is distortion caused by implanting false memories into users' long-term memory through leading or suggesting questions. This could be avoided by providing users with appropriate questionnaires.

5.3.3.3 Long-Term Memory Retrieval

Information is extremely organised inside our brains in the long-term memory like a dictionary (Gray, 2001). Perhaps the means by which we retrieve information from it (search within the dictionary) is the most fascinating part of the whole process. Mental associations are usually carried out in order to interrogate our long-term memory, like thinking about the colour "red" and remembering "strawberries", "apples", "London Buses" as a result. These are called retrieval cues, and they help us probe our memory to find appropriate items to any category we think of.

Aristotle, more than 2000 years ago, proposed the law of association by contiguity, that is, if a person experiences two environmental events (stimuli) at the same time, or one after the other (contiguously), those events will be associated in the person's mind. And in turn, the thought of one will elicit the thought of the other (Horthersall, 1995). This could be useful in our specific scenario by presenting the users to view a virtual environment of their route, and

that would elicit the thought of things like when they got off the bus, parked their car, or which turn they took. This is also enforced by another concept of retrieval called context-dependant memory, where recall depends on similarity between the test environment and the original encoding environment.

5.3.3.4 Spatial Cognition

An important element to consider in this section is the spatial nature of information being dealt with. In order to account for different users' abilities to deal with space, we need to look closer into understanding different humans' spatial cognition. Spatial Cognition is concerned with the acquisition, organization, utilization, and revision of knowledge about spatial environments. These capabilities enable humans to manage basic and high-level cognitive tasks in everyday life.

Montello (2001) describes cognitive systems as including sensation and perception, thinking, imagery, memory, learning, language, reasoning, and problem solving. On the other hand, spatial properties include location, size, distance, direction, separation and connection, shape, pattern, and movement. It is suggested that spatial knowledge of places is developed in a sequence of three stages as follows.

1. **Landmark Knowledge:** Unique features that identify a place.
2. **Route Knowledge:** Travel routines that connect ordered sequence of landmarks.
3. **Survey Knowledge:** Corresponds to map-like mental knowledge. This could be enhanced by the exposure to maps of places, which is becoming very common nowadays with the availability of smart phones with many people (e.g. iPhone, Android... etc.).

People tend to store this spatial information as a cognitive (mental) map. This metaphor was introduced by Tolman in a (1948) paper to refer to a mental representation of spatial layout of the environment. The spatial information stored includes the landmarks, route connections, as well as non-spatial attributes (e.g. "the road where I rode my first bike"). Therefore, in the application presented here, we make use of Google Maps as the background for the interface which provides users with landmark locations along with their tracks. And even if users don't have good survey knowledge, they still can use the Street View background that provides photos of the route, which help invoking the retrieval cues to recall the exact route taken.

As stated in the previous subsection, there are many advantages of having GPS travel diaries with map-based recall surveys. A definite advantage is due to the fact that when humans communicate space via language, they tend to express mostly non-quantitative or imprecise "fuzzy" quantitative information. This is due to the lack of survey knowledge among most of people. And therefore, using map-based recall surveys are highly advantageous.

Another advantage with map-based recall surveys is that association by contiguity is achieved by presenting the user with the real map that evokes his/her mental map. As a result, association is achieved and hence retrieval cues evoke the memory to remember the user's state at different parts of the journey.

Other elements to account for are the similarities and differences between different groups. Some people are better at tasks such as way-finding, learning spatial layouts, or reading maps. There are many factors that may contribute to these variations in spatial cognition: body size, age, education, expertise, gender, social status, language, culture, and more. This creates uncertainty in the level of accuracy of the reported information. Therefore, the application builds a profile for each user compiling all their personal information to be able to draw associations.

5.3.4 Geo-Web 2.0

Nowadays, with the eagerness of people to take part in generating web content, concepts like crowd sourcing came into existence, which demonstrated the internet users' enthusiasm to contribute content (Howe, 2006). Web users feel they can express and post their version of interest of more people about where everything is. This has matured people geographically and grew their location awareness.

5.3.4.1 GIS Users to GIS Technicians

Moreover, the availability of programming platforms like the Google maps and Google earth APIs that used very useful emerging technologies like AJAX, has led to the concept of "Democratisation of GIS" (Goodchild, 2007). This also led more users to understand basic concepts of GIS, and even turned some into GIS specialists. An obvious example is OpenStreetMap (OSM) where users contribute data for producing maps of the world online (Hacklay & Weber, 2008). This has encouraged and motivated people to collect their own GPS data and post it online. Goodchild (2007) argues that the reasons behind this motivation could be self-promotion or the satisfaction that users derive from knowing that other people would be interested in seeing and even using their data. Other users are motivated by making information available to friend and relations. In this section, we attempt to take advantage of this motivation and make use of it to replace the traditional travel diaries. We hope that this would pave the platform for bridging this link between the online Geoweb 2.0 capabilities and the previous horrendous process of collecting user transport data.

5.3.4.2 Online Applications

To the knowledge of the author, no application was found that serves this specific purpose. However, there are a number of similar applications that have different purposes and accordingly have different functionalities to serve that purpose. The mapmyfitness.com application enables users to upload the GPS tracks and states which type of sport they were undertaking. It returns analysis on their workout along with a representation of their tracks on a Google Maps background. maps.inersource.com also enables users to view their GPS tracks, routes and way points on different maps background for sharing and social interaction. Also, a similar interface is everytrail.com for sharing travelling experiences around the world on the web. Very few applications provide the functionality of editing GPS tracks online from a map interface such as gpsvisualizer.com. Also, some desktop applications enable users to perform edit functionality (Viking software), but yet not having the functionality of adding any attributes to the data. In this section, we aim to have transport related edit functionalities to fill the missing gap in providing users with more transport-based GIS web functionalities.

5.3.5 Intelligent Travel Diaries

The system developed in this study is an attempt to regulate and automate the travel diary concept in a Geoweb context. The application is also published in Bolbol et al (2010). The system focuses on attaining the most adequate form of GPS data, in order to feed adequately into the travel activity inference model and obtain the best results. The system consists of multi-layered processes that flow in a waterfall manner from one process feeding into the following one.

5.3.5.1 Motivation: Aim of Application

After glancing through the literature, it becomes obvious that different phases of travel activity inference methods so far are unregulated. Also, there appears to be a lack of a systemized and guide-lined process of data collection for that purpose. Therefore, there appears to be an urging need for developing a system that is a natural step forward in the way travel diaries and follow-up surveys are collected and maintained. The following are the main aims of the application that make it an advantageous upgrade of the traditional travel diaries.

Online: A major advantage is having the system in an online environment which takes the burden off the users of saving their tracks and managing them.

Interoperability: Using a standard format of GPS files (GPX), which makes the input and output of the system flexible to read files from other GPS sources, and to be able to use the system's data on other platforms.

Visualization: Having Google Maps and Google Earth in a 2d and 3d environments respectively as a background map, enables users to have a better recall of their trip details. Another useful functionality in that context is using Google StreetView for viewing the trails, which simulates a realistic view of where users have been, which has in turn a major effect on their retrieval cues, where retrieval cues can provoke powerful recollections. This has led some researchers to speculate that there is a high possibility that all memories are permanent. The way to bring back these memories is by triggering some elements that were in the original environment of the original event when it took place. These elements are the retrieval cues, and having images (StreetView) simulating the actual route users took in the original event acts as some of these elements.

Trip Recall Time Gap: As mentioned in subsection 5.3.3, time is a major factor for memory loss. However, the rate of loss is very gradual after the first short period. Also, real time applications have numerous disadvantages like the ones mentioned in subsection 5.3.2. Among these is the problem resulting from users feeling a privacy invasion as a result of the real-time notion. This trade-off is illustrated in Figure 5.8 as muffin-top-shaped curves. Therefore, this application gives the necessary time gap that the user needs to feel more at ease of the "reporting my location" notion. It also provides users with a good incentive by capturing their enthusiasm to upload their cycling, walking or driving tracks, which makes them eager to upload them regularly.

User Burden: The application also imposes fewer burdens on users by not having to input their metadata (or label their data) as opposed to mobile phone trackers.

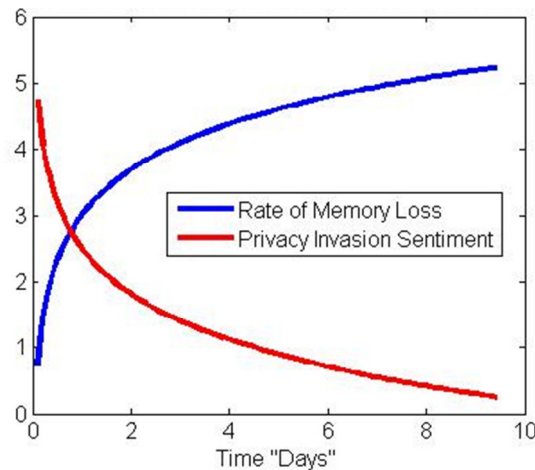


Figure 5.8 Trade-Off between Rate of Memory Loss & Privacy Invasion Sentiment

User Interface: The interface also enables the users to change and input data visually from the map itself which makes the tasks required by the user easier, faster, and more accurate.

Incentives: As illustrated in subsection 5.3.2.3, participants need incentives in order not only to add data, but to spend time adding information to their collected data. The application therefore attempts to take advantage of the incentives behind user generated content the in Geoweb, as described in subsection 5.3.4.1. The interface allows users to run track analysis; so as cyclists and pedestrians enjoy watching their tracks and performances. It also gives feedback on the “green travel” ratio of the users’ trips, which also generates a sense of awareness about people’s travel habits and their impact on the environment. The interface also allows users such as holiday travellers to post their tracks in a shared community on the website adapting the idea of volunteered Geography (Goodchild, 2007) invoking self-promotion and sharing their experiences with friends.

5.3.5.2 Data Collection - GPS Devices

The application currently only supports data from GPS Devices of GPX format. The devices used in testing this application are developed by u-blox (described in chapter 4). These devices have a very reliable power saving property which enables the users to collect data for till more than 3 months. They also have accelerometers that switch the device on only if motion is detected, which increases the battery life and memory storage space very effectively. The devices are quite small in size too, which makes them attractive to use in tracking applications where they could fit unnoticeably in a jacket pocket or bag. Other file formats from other devices are planned to be also accepted for upload in the future like Garmin GPX, KML. Users will also be able define their own XML formats in the future to be used in the type of files they upload. Also, plans in the future are in place to use formats from other devices that use different positioning sensors.

5.3.5.3 System Architecture

The conceptual design of the system is broken down into three phases/segments (as illustrated in Figure 5.9). The first segment is the data collection where all tracks are collected

whilst the users' trips using the GPS devices (or other sensors). The second segment is where the GPS processing happens using the ephemeris data. In the case of this test using the u-blox devices, the processing happens using a software package (YUMA) enabling a map interface to display the processed tracks. YUMA provides the output data in three formats: KMZ, CSV, and GPX. The third segment is where the main user interaction happens in an online environment. We call the application the GeoTravelDiary (accessible from geotraveldiary.com). The application takes into account of all the specified advantages described in subsection 5.3.5.1.

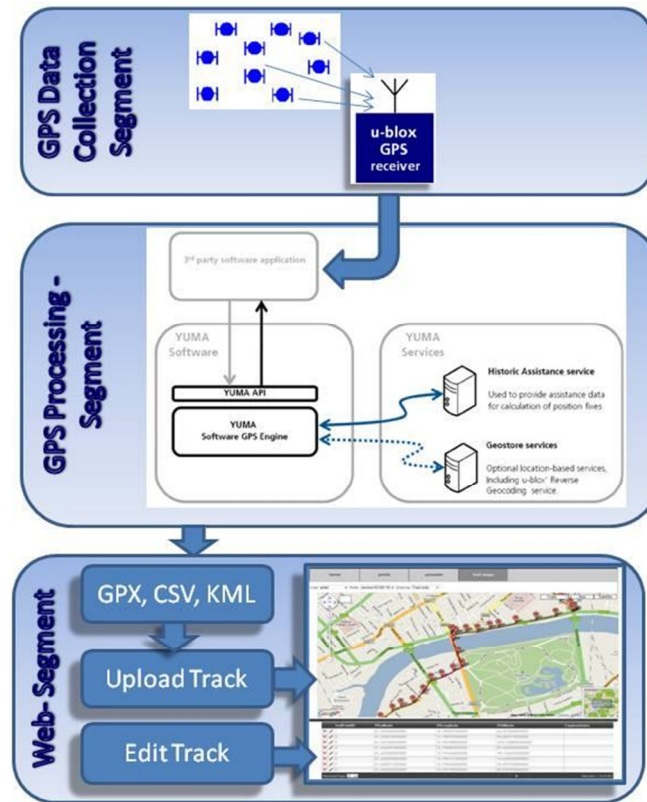


Figure 5.9 The System's Data Flow

5.3.5.4 The interface (www.geotraveldiary.com)

Technology Used

A .NET3.5 framework is used along with a SQL Server 2008 Database, while the front end is developed in ASP.NET using XHTML, CSS, AJAX, JavaScript and JQuery. Object-oriented design is used and N-Tiered Architecture using Microsoft Enterprise Library 4.1 with static queries in stored procedures. Unlike normal web sites and applications, this site is developed using a post-back free/single page interface. Once the user logs in, they are taken to the Main page after which they remain on that one page till they log out. The menus on that page load new controls on the same page using AJAX, JavaScript and JQuery. This gives the user a much better response time to clicks and reducing network traffic by avoiding the transfer of the same content repeatedly. This sort of an interface, when required on the internet, is normally developed by JAVA embedding or using Flash/Silverlight or similar technologies. Each one of these technologies requires some sort of installation on the client side besides other issues like difficulty in integrating with web services and databases. Although the application has data

driven content and menus, it has various hard coded features put in to speed up loading times. The application also arranges the storage of state information in a way to maximise efficiency.

The application is very scalable and the design allows rapid development. Free use of JavaScript is avoided and it is mostly done in tight bundles of pre-tested code packed together in controls. This allows the developer to build on the application and access JavaScript functionality using C#. This makes the application easy to debug during development and testing thus making upgrades quicker and cheaper.

Data Flow

The data for the tracks is obtained by the users uploading their individual tracks on to the system. Once received, the data is parsed and stored as points, data relating to the points and relationships between the points on an online SQL Server database. Other useful information about the satellites is also saved in the RDBMS in appropriate tables.

We are able to run spatial queries directly on the database by using the special spatial features available in SQL Management Studio. The data is also retrievable into the ASP.NET application via Stored Procedures. In the application, the data is shown on its own, in grids, used to compute figures or to build graphs even as points and lines on maps. The data for the individual points and relationships between them, obtained from the database are used to display markers and to connect them together to show approximate tracks.

In the application, the data is used by grouping it together as objects to mimic the structure of the database. This allows easy manipulation of the data while still maintaining the integrity of the database. Since the data access logic is embedded into the lower tiers of the application, a front end developer can easily add more functionality to the application without having to worry about the complex data relationships.

UI-Interaction

The main aim of the application is to mainly interact with the interface through the map. This enables taking more advantage of visualization. However, users could also do administrative tasks like changing their personal details through normal page interaction. Also, users will have to upload their tracks beforehand in CSV, KMZ, KML, or GPX formats.

The map background could be viewed in several flavours (Map, Satellite, Hybrid, Earth (3d), or Street View). Also, having the option of viewing Street View along with the Map enhances the user's experience to be able to link the tracks' location on the map with the images of the route taken.

The edit capabilities of the user could be classified into two types: Marker Edit or Data Edit. Marker Edits are edits to the Google markers on the map. These functionalities are such as dragging, deleting or adding the markers (see example in Figure 5.10). Data Edits are edits done to the attributes of the points of within track. This includes editing information about the transport mode and type of location (work, home... etc.).

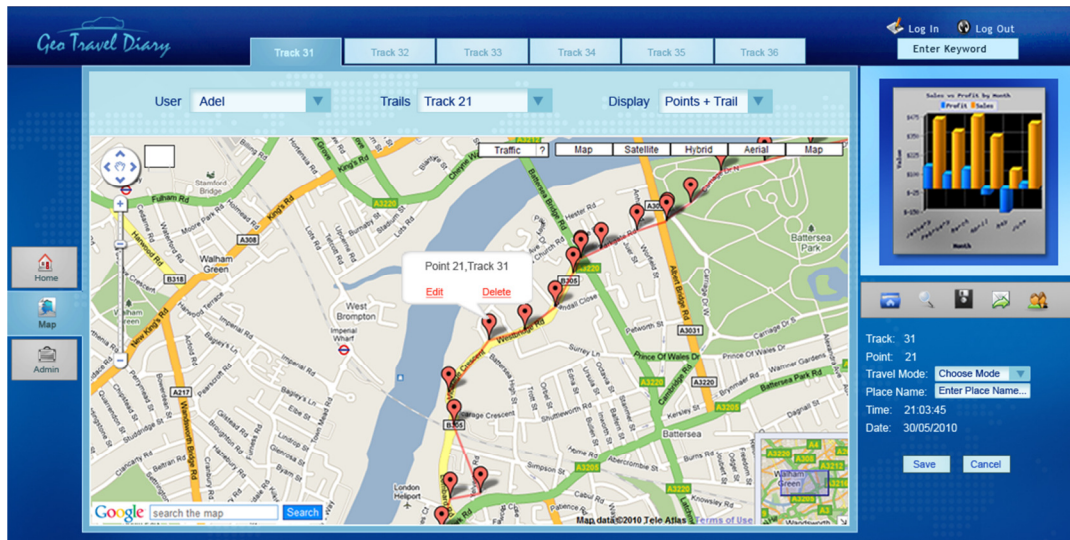


Figure 5.10 The Application's Interface Showing Point Edit Functions

Once a transport mode is selected for a point in the trip, all the points that follow are assigned the same mode until a new mode is selected for any point afterwards. Also, the colour of marker changes according to the transport mode, which gives an indicative visualization of when the users used different modes. Users can also select to view their tracks by date, mode or user. Moreover, users can run analysis of what percentage of their travel was environmentally friendly or an average of how much calories they burned in a week, or even how many miles they drove their cars. These types of analysis are considered as an incentive to motivate users to accurately input their attribute information such as travel mode changes.

5.3.6 Application Conclusions & Further Work

Introducing such a system in an online environment tackling all the usability, response incentives, and visualization issues clearly has the potential to change the way travel diaries and travel information has been collected in the past. This has been shown where the application overcomes the main disadvantages of traditional travel diaries such as speed, cost, and response rates.

The introduction of a map-based interface where all the user edits are inputted makes non-GIS users perform GIS view and edit functionalities very easily. And the fact that it is in an online environment aids the objective of providing users with incentives such as web-contribution, emissions awareness and sharing experiences with other users. This also makes the process fast, easy to share, and useful for having a large database for running travel and traffic analysis. The application also provides the tools that provoke retrieval cues in the user's long-term memory by using background mapping along with Google StreetView tool which acts as a strong association tool.

Further work can extend the current application regarding sharing trips on various social networks (such as Facebook) to provide large scale awareness of transport issues and alternatives. Also, more functionality is to be added to the map interface in order to give the users more flexibility to add their information along with a better testing with a large number

of users. Also, a second phase is to use the collected labelled data to develop an algorithm that learns and deduces the travel information automatically.

5.3.7 Application Testing & Implementation

As described in the previous chapter, the data collection process involved recruiting a pilot group of 21 participants and then a main validation group of 84 participants. The process was similar for both groups with some minor differences that will be described in the text of this subsection. The process mainly consisted of 4 phases; namely, the placement interview, testing period, exit interview and possible follow-up interviews.

5.3.7.1 Placement Interview

The participants were first invited for a placement interview that often took place at UCL's headquarters with the exception of some cases. The interview first consisted of a general explanation of the experiment, its purpose and the intended outcomes from running it. The participants were then handed the GPS tracking device and were explained how to use and maintain it. They were also handed a crib sheet explaining the significance of different signals the device produced. It is worth to note that many participants joined the experiment as a result of their genuine interest in the topic, and therefore, this orientation session usually extended into a prolonged conversation. The participants were then asked to sign an experiment consent form and to fill out the placement interview questionnaire (Appendix A) which contained personal information, details of typical places of interest to the participants and transportation modes used to commute to and from these places. The participants were also handed a Paper Travel Diary Log (Appendix C) to optionally fill with details of their trips during the experiment period highlighting that filling them would provide the experiment with richer data.

5.3.7.2 Data Collection Phase

During the eight weeks of the GPS data collection of the validation group (84 participants) and two weeks of the pilot group (20 participants), the participants were instructed to contact the experiment conductors in case any alerting signals occurred during the experiment period. There were several device failures during the experiment and device swaps were arranged as necessary. An appointment was typically arranged for this purpose and the experiment conductor would make a trip to the participant do the swap.

5.3.7.3 Exit interview

Finally after the data collection period, the participant would be invited for a second time to an exit interview mainly for the purpose of gathering feedback on the experiment's process and any exceptional occurrences. The interview started off by having a general conversation about the process and any oddities that occurred with regards to the device. Discussions also took place with regards to filling the Travel Diary Log or unusual activities that the participant has made during the collection period. The participants were then asked to fill in an exit questionnaire (Annex B) to document and measure their feedback on different experiment aspects.

On some occasions, the participants were also presented with their tracks on the GeoTravelDiary application containing their tracks automatically segmented into walk/non-walk stages using a basic version of part of the developed framework (to be explained in detail in chapter 7). The participants were observed during this exercise for interaction with the interface and were also invited to ask/discuss their thoughts with the experimenter. As mentioned, this was an occasional occurrence as the data was not always ready in form to be presented on the interface, so a further appointment was arranged for these other cases.

5.3.7.4 General feedback

The vast majority of the experiment participants have not fully filled their paper Travel Diary Log. An average of around 21 trips was usually recorded on these sheets mostly covering the first few days of the device carrying period (if any). An average scoring of 8 on the scale of 1 to 10 (with 10 the highest) was calculated for responses regarding the level of burden to fill the diary. When verbally asked during the exit interview about the paper diary process, many comments mentioned that it was an extra worry they had to care about and usually had little time for.

The developed intelligent travel diary (GeoTravelDiary online application) was only rolled out for the pilot participants to use during the collection period, but was tested on the validation participants only after the experiment had ended. This was due to existing bugs that made the interface slightly vulnerable to use by several user end points simultaneously. The online application was reported by the pilot group to be “easy” to use and “interesting” to view one own tracks daily as they uploaded them online. Another worry was that the pilot participants have reported to be more wary about their daily movements as they always remembered that they were tracked since they were constantly reminded by the online application. That made it imperative to take exposure to the application out of the experiment’s process for the validation group as a cautious measure.

As previously mentioned, on the exit interview day (or on another day after the exit interview), the participants were asked to check their tracks on the online application while being observed by an experimenter. The participants were provided with their tracks automatically segmented into walk/non-walk stages using a tool developed as part of the framework produced by this research. On many occasions, participants did not require assistance with the application after they were provided a brief demonstration on how to use it. In most cases, the stage splits were reported to be performed very efficiently, however, several occasions occurred when two consecutive stages of similar modes required to be merged together. A recommendation out of this exercise is that presenting the participants with such an interface is best after the testing period for the reasons discussed above. Another recommendation is that the participants are best to be observed and talked to while providing their activity feedback on the application in order to refresh their long-term memory through discussion on specific events that occurred on odd occasions in the experiment period.

Other feedback from the exit interview questionnaire included levels of burden and problems faced by the participants regarding the use of the devices and the experiment as a whole. When asked how easy did participants find it to charge the device daily, an average of 2.7 was

scored on a scale of 1 to 5 (5 being very difficult) and an average of 2.1 was scored on a similar scale in response to how easy it was to remember to carry it. Also, 90% of participants have reported that the device has run out of charge at least once during the experiment period, and a similar figure was reported for forgetting to carry the device at least once during that period. Reasons for forgetting the device ranged from changing coats or bags to forgetting it on the charger. This reflects the significant yet bearable burden that participants have had in order to manage the device. A recommendation here would be to use a device that has a longer battery life. Moreover, with the rise of wearable GPS devices nowadays such as GPS watches; a wearable technology would significantly make minimise the burden of remembering to carry the device.

5.4 Summary of Data Pre-Processing Strategy

In this chapter, we introduced the guidelines to pre-process GPS data coming from travel surveys making use and building up on practices in previous literature. This pre-processing phase involves issues such as data cleansing, choice of modes, data generalisation and data labelling. The chapter addresses each of these issues separately leading an attempt to standardise the data pre-processing phase which leads to unifying the basis on which the data is structured and labelled. This standard unification assures that the uniform quality of the metadata for datasets from different participants and different situations, and hence, standardising the validation method of the classification algorithm developed for this research.

This chapter first starts by addressing the GPS data accuracy limitations. We take advantage of some data cleansing methods such data filtering and identifying stops in order to overcome low positional GPS accuracy problems. As for signal loss problems, we use filtering techniques described at the beginning of the chapter to filter out partial blockage using HDOP, satellite, altitude and speed information of every GPS fix. We also use total blockage to identify trip ends and we further describe it as part of chapter 7. We also address the modes of transport to be studied in this research to include a wide range of transport means that London possesses. These modes are: (1) bus, (2) cycle, (3) car, (4) train, (5) underground and (6) walk, as well as stops in a GPS track. This wide variety of transport modes to test the developed classification model enriches the medium in which the validation will take place. This variety would also reflect how robust the developed method is to deal with different transport modes.

The chapter then describes the different levels of detail that might exist in GPS track information (or metadata). We define the level of information granularity by defining these levels of a track being: (1) Trip, (2) Stage, (3) Segment and (4) fix levels. We also specify clearly how each level is defined highlighting what defines stops. These clear definitions are aimed to standardise the process of attaching metadata (such as the mode of transport) for participants when providing feedback on their tracks. For that purpose, we also define the decisions to be made whilst labelling the tracks in specific confusing situations such as parking a car or sitting outdoors. This standardisation process unifies and removes the bias from the metadata (transport mode) by which the classification method developed in this research is based on when tested and validated.

This chapter then describes the medium through which the participants of this study shall label their tracks providing the transport modes they used and stops committed. We develop an online web-based system where participants add such information to a web-based GIS interface. The interface emulates traditional GPS-based travel diaries overcoming many limitations that current travel diaries possess. These limitations include the difficulty of recalling spatial information or low response rates and inaccurate information due to issues such as participant privacy, assigning burdensome tasks, battery constraints and the lack of incentives. Providing such a medium that encourages participants to accurately provide feedback on their tracks assures that the basis on which the building and validation of the developed classification algorithm are accurate, and hence, producing better quality results.

The final part of the chapter describes details of the experiment of data collection process that involved recruiting a pilot group of 21 participants followed by the main validation group of 84 participants. A summarised description of feedback on the task load of dealing with the GPS device and the paper Travel Diary Log was also mentioned to provide a full understanding of the level of burden the participants endured with regards to different tasks. The feedback gathered by different meet-up checkpoints along the experiment process reflected several important findings. A high average level of burden was scored for maintaining the paper Travel Diary Log, which reflected the need for some form of intelligent travel diary application and enforced the requirement for this research as a whole. Among other recommendations resulting from this experiment is exposing the participants to the GeoTravelDiary application right after the data collection period rather than during it. The brainstorm process while interacting with this application has also been noted to refresh the participant's memory through story-telling (from part of the participant). A final recommendation was to use tracking devices with a longer battery life in order to decrease the burden level of device-related tasks.

Chapter 6

Phase I: MODE CLASSIFICATION (MOVING WINDOW- BASED SUPPORT VECTOR CLASSIFICATION (SVC))

6 PHASE I: MODE CLASSIFICATION (MOVING WINDOW-BASED SUPPORT VECTOR CLASSIFICATION (SVC))¹⁰

6.1 Introduction

As first described in chapter 3, the mode of transport inference framework from sparse GPS data consists of three phases; namely, classification, segmentation, and network matching. This chapter describes the classification phase as the first of the three constituting this framework. First, unlike many previous attempts, we carry out an ANOVA analysis to identify the Independent Variables (IV) that best discriminate between different modes from GPS data tracks. Once the IVs are identified, we use them as inputs into the classification framework we develop in this chapter. The classification framework we develop is based on Support Vector Machines (SVM) technique, overcoming some of the limitations other techniques possess when dealing with GPS data. Moreover, the classification uses a moving window along a GPS track to classify parts of the track strings collectively rather than individually. This collectiveness allows understanding patterns of movement that each mode is characterised by, and hence, enhancing the achieved accuracy. An initial modal segmentation process then is implemented to reason between modal switches using a transition matrix we compile based on the data available to this research. Finally, we discuss the classification results and their accuracies at the end of this chapter, highlighting the areas where the rest of the inference framework needs to address. These areas are then addressed using the segmentation and network matching phases described in chapters 7 and 8 respectively.

As discussed in chapter 2, a diverse account of work has emerged in the previous decade attempting to infer the transportation mode from GPS data. This inference could largely replace or complete a lot of the feedback required by users when labelling and tagging travel diaries. Studies aiming at inferring the transportation mode could be divided into procedural and Machine Learning (ML) approaches. Procedural approaches attempt mainly to make inferences based on logical assumptions, such as how a typical person would travel (Stopher, et al., 2008a). Other assumptions include the surrounding environment, such as the nearest transportation networks (Chung & Shalaby, 2005), or; the temporal logic assumptions of activities, such as people are more likely to have no activity after mid-night (Liao, et al., 2007). On the other hand, ML approaches attempt to do the inference based on learning from existing data, possibly combined with similar logical assumptions. Examples of these studies use Decision Trees (Zheng, et al., 2010; Reddy, et al., 2010; Manzoni, et al., 2010), Bayesian Networks (Stenneth, et al., 2011), Fuzzy Logic (Schüssler & Axhausen, 2009), Hierarchical Conditional Random Fields (Liao, et al., 2007), and Support Vector Machines (SVM) (Zheng, et al., 2008). These ML approaches could be broken down into two aspects; namely, *selection of variables* for inference (or combination of variables), and the *method of inference* (the details of the learning algorithm used).

First, there are several issues regarding the *variable selection* to be used for the inference. Different studies use different variables (or combination of variables), such as speed,

¹⁰ Most of this chapter is based on a publication of ours:

Bolbol, A, Cheng, T, Tsapakis, I & Haworth, J, 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification, *Computers, Environment and Urban Systems*, Special Issue: Advances in Geocomputation, Volume 36, Issue 6, pp. 526–537.

acceleration, maximum or medians speed, and acceleration and length between GPS fixes. However, none of the studies, to the best knowledge of the authors, have based their variable selection process on statistical evidence. Second, the technique-related issues include the usage of a limited number of transportation modes in the learning, the high dependence on segmentation into transportation modes, and high reliance on temporal information. Moreover, some technique-related assumptions are often made in previous work such as that certain modes cannot follow each other in a GPS sequence and that every two GPS consecutive fixes are analysed individually not accounting for the track as a whole. Therefore, we attempt to fully understand and account for these aspects in the process of classification. We aim to solve a classification problem of GPS data into different transportation modes (*car, walk, cycle, underground, train and bus*).

In this chapter, we describe how we address the variable selection and method-related issues. Before introducing the classification framework, it is essential to select the best classifier(s) or independent variable(s) IVs to be used to classify GPS points into transportation modes (Mitchell, 1997). Therefore, we first run an analysis of variance (ANOVA) test to select the IV(s) that best discriminate between the different transportation modes. In turn, this should improve the performance of whichever classification algorithm that would be used in the following phase. We statistically compare the candidate variables using different statistical measures, such as Wilks' Lambda and between-groups F to assess each variable's discriminatory power. The results from the classification, based on the selected variables are then analysed and compared illustrating the power of each over different modes (categories). This analysis is presented in section 6.2.

Finally, we attempt to identify transportation modes from the collected sparse GPS data, without information or assumptions about the participant's temporal or location contexts, which some of the previous approaches were based on. We attempt to apply the same inference method later in chapter 9 with the knowledge of some information about the participants such as the ownership of a bike or car in order to enhance the accuracy. For the classification, we use Support Vector Machines (SVM) to perform the inference from speed and acceleration values calculated from GPS data. Due to its high quality of out-of-sample generalization and ease of training, SVMs provide far beyond the capacities of traditional ML methods used in previous research which are discussed in section 6.3. Furthermore, using SVMs, the selected kernel could be applied directly to the data without the need for a feature extraction process. This is advantageous in the context of learning from the structure of the data, since a lot of this structure is lost by the feature extraction process (discussed in detail in section 6.3.2). This enables us to study a sequence of movements of a participant rather than each movement individually, and hence, achieving a better classification. We achieve this by using a moving window that classifies instances of data consequent blocks (section 6.4). We complement this by using logical filters that apply a transition matrix between different phases of the trip (section 6.5). The results of this inference are presented in section 6.6 along with some discussions and conclusions in section 7.1.

6.2 Independent Variable Selection

Generally in a classification problem, the variable that is to be predicted is known as the dependent variable (transportation mode in our case) because its value depends upon, or is

decided by, the values of all the other attributes. The other attributes that help to predict the value of the dependent variable, are known as the independent variables (IVs) in the dataset. The less correlated (or statistically dependent) the IVs are the less the outcome of the classification is inclined to be biased.

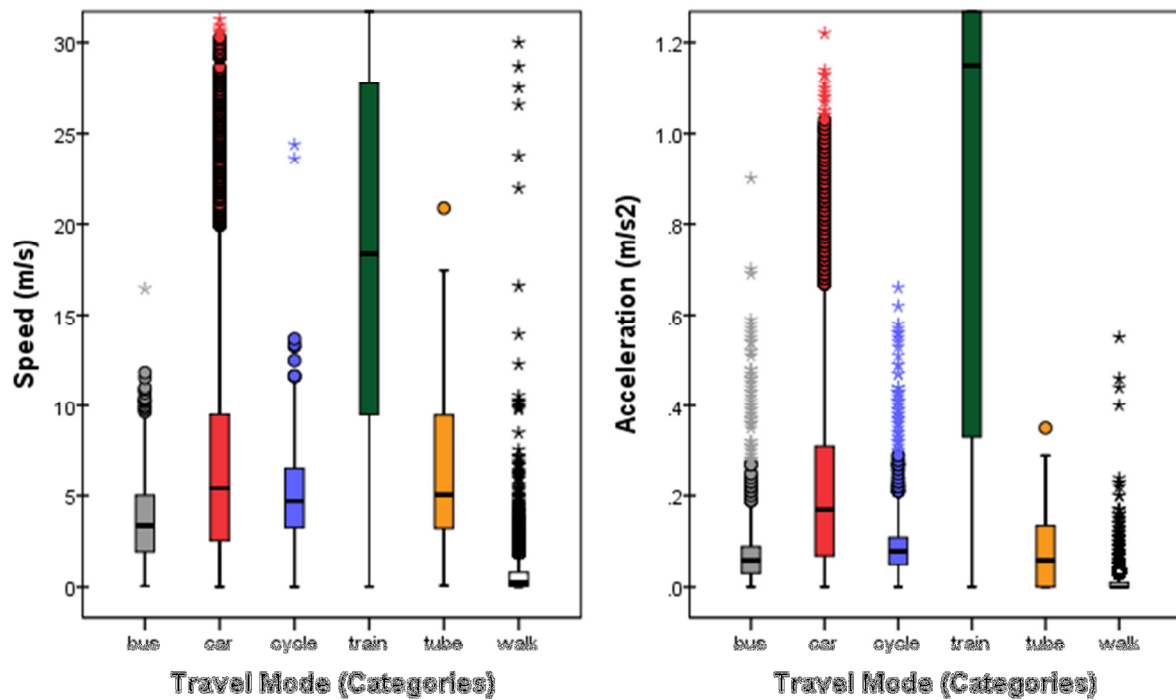
A major limitation in methods attempting to infer the transportation mode is the choice of IVs to be used for classification. For most studies the variables chosen were not based on any statistical evaluation justifying the variable choice being made. Most studies use variables such as length, speed, acceleration, maximum or median of speed or acceleration through a stage (Schüssler & Axhausen, 2009; Zheng, et al., 2008), either together or alone for classification without providing a statistical basis for the choice. The correlation of the chosen IVs in these studies was neither accounted for. Therefore, in this section we conduct a statistical evaluation of different IVs that could discriminate between different classes (modes) in this classification problem. The outcome of the evaluation identifies the best IVs to be used for the classification.

6.2.1 ANOVA Test for Variable Selection

Four potential variables were taken into consideration for the analysis; three of which are distance, speed and acceleration, which are highly inter-correlated where they all stem from one another. We also consider the change rate in heading (direction) as was suggested by a previous study (Stopher, 2008).

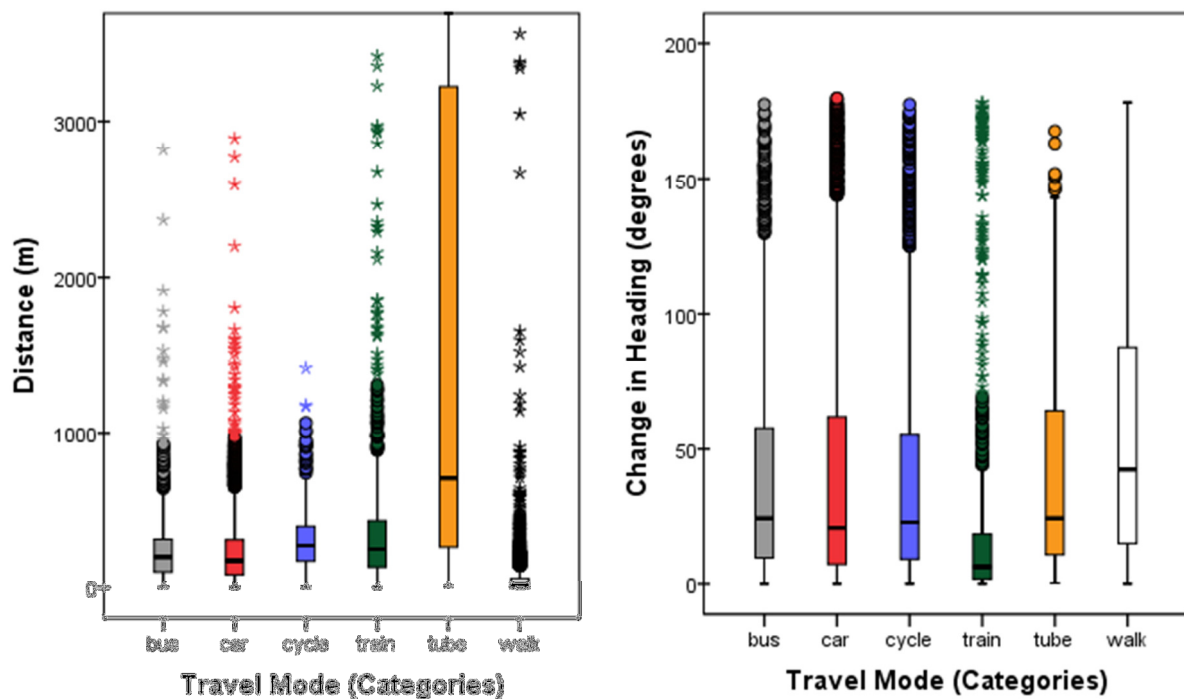
The testing sequence starts with group statistics to examine the differences between the categories on each of the independent variables using category means and ANOVA test. The mean differences between distance, acceleration and speed suggest that these may be good discriminators as the separations are large. This separation is clear in Figure 6.1 representing the distribution across different modes as box plots for each variable in a separate plot. These 3 variables effectively discriminate the *walk* and *train* modes from the rest, as illustrated in the figures. However, acceleration appears to discriminate the *car* mode from the rest quite well. On the other hand, the rate of change in direction does not seem to significantly discriminate between any of the modes, except for the *train*, which could be caused by the fact that *train* trajectories follow fixed tracks for long distances (Figure 6.1d).

6 Phase I: Classification (Moving Window-Based Support Vector Classification (SVC))



(a) Speed plot for different transportation modes

(b) Acceleration plot for different transportation modes



(c) Distance plot for different transportation modes

(d) Change in heading plot for different transportation modes

Figure 6.1 Box Plots for the Values of different Independent Variables

Equality of group means results are presented in Table 6.1. In order to assess the discriminability of the different IVs two statistical measures are introduced: the Wilks' Lambda Λ and the Between-Groups F . The former is used in multivariate analysis of variance

(MANOVA) to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables (Everitt & Dunn, 2010). Wilk's Lambda is a statistic that takes into consideration both the differences between groups and the cohesiveness or homogeneity within groups (Klecka, 1980). However, a variable which increases cohesiveness without changing the separation between centroids may be selected over a variable that increases separation without changing the cohesiveness. When the IVs are

considered individually, Λ is given from

$$\Lambda = \frac{\text{Within Groups Sums of Squares}}{\text{Total Sums of Squares}} = \frac{w_{il}}{t_{il}}$$

Equation 6.1.

$$\Lambda = \frac{\text{Within Groups Sums of Squares}}{\text{Total Sums of Squares}} = \frac{w_{il}}{t_{il}}$$

Equation 6.1

$$w_{il} = \sum_{j=1}^g \sum_{k=1}^{m_j} f_{jk} X_{ijk} X_{ljk} - \frac{\sum_{j=1}^g \left(\sum_{k=1}^{m_j} f_{jk} X_{ijk} \right) \left(\sum_{k=1}^{m_j} f_{jk} X_{ljk} \right)}{n_j}$$

Equation 6.2

$$t_{il} = \sum_{j=1}^g \sum_{k=1}^{m_j} f_{jk} X_{ijk} X_{ljk} - \frac{\left(\sum_{j=1}^g \sum_{k=1}^{m_j} f_{jk} X_{ijk} \right) \left(\sum_{j=1}^g \sum_{k=1}^{m_j} f_{jk} X_{ljk} \right)}{n}$$

Equation 6.3

where:

- g = number of groups,
- p = number of variables,
- i, l = 1, ..., p ,
- X_{ijk} = value of variable i for case k in group j ,
- X_{ljk} = value of variable l for case k in group j ,
- f_{jk} = case weights for case k in group j ,
- n_j = sum of case weights in group j ,
- n = total sum of weights,
- m_j = number of cases in group j .

In Table 6.1, large values of lambda indicate that group means are close, while small values are indicators of different means. Acceleration and speed seem to be the best discriminators in this case, with a small difference between their performances.

| | Wilks' Lambda | F | df1 | df2 | Significance |
|----------------------------------|---------------|----------|-----|-------|--------------|
| Distance (m) | 0.896 | 441.820 | 5 | 18934 | 0.000 |
| Speed (m/s) | 0.486 | 4004.532 | 5 | 18934 | 0.000 |
| Acceleration (m/s ²) | 0.459 | 4462.582 | 5 | 18934 | 0.000 |
| Difference in Heading (Deg) | 0.965 | 135.945 | 5 | 18934 | 0.000 |

Table 6.1 Tests of Equality of Group Means

The second statistical measure used, the Between-Groups F , takes into consideration the sample size of the groups. This differs from a test that is solely based on squared distance (Klecka, 1980). Comparisons between small groups will be given less weight than comparisons between large groups. The advantage here is that this criterion will maximize differences between pairs containing larger groups. Acceleration and speed are still the best discriminators in this case; however, the difference between them is higher. This finding could be attributed to the sample size of the *car* mode having a high significance in manipulating the value of this statistical measure. The following equation 6.5 is used to calculate the F statistic based on another statistic called Mahalanobis Distance D^2 , which is the distance between two groups (a and b) (Klecka, 1980) and is calculated from equation 6.5.

$$F = \frac{(n-1-p)n_1 n_2}{p(n-2)(n_1 + n_2)} D_{AB}^2$$

Equation 6.4

$$D_{ab}^2 = (n-g) \sum_{i=1}^p \sum_{j=1}^p w_{ij}^* (\bar{X}_{ia} - \bar{X}_{ib}) \times (\bar{X}_{ja} - \bar{X}_{jb})$$

Equation 6.5

where:

- n_z = sample size of the group z ,
- \bar{X}_{ia} = mean of i^{th} variable in group a ,
- \bar{X}_{ja} = mean of j^{th} variable in group a ,
- \bar{X}_{ib} = mean of i^{th} variable in group b , and
- \bar{X}_{jb} = mean of j^{th} variable in group b .

It could be noted that speed appears to be a better discriminator for only some categories when calculating the Wilk's Lambda and the Between-Groups F . On the other hand, acceleration is better for most of the categories and/or for the categories of the highest sample sizes. From Figure 6.1c we could also note that the *car* mode is better discriminated using acceleration rather than speed in Figure 6.1a. We therefore ran the same analysis again but this time using only 3 categories namely: *car*, *train* and the rest of modes aggregated into one category. This categorisation was due to the natural division of the acceleration data illustrated in Figure 6.1c. Table 6.2 shows the results of this second run, proving that acceleration produces a better discrimination of the 2 categories. It also performs much better than speed, while it yields a bigger difference than that shown in Table 6.1.

| | Wilks' Lambda | F | df1 | df2 | Significance |
|----------------------------------|---------------|-----------|-----|-------|--------------|
| Distance (m) | .970 | 293.326 | 2 | 18934 | 0.000 |
| Speed (m/s) | .540 | 8079.834 | 2 | 18934 | 0.000 |
| Acceleration (m/s ²) | .464 | 10951.800 | 2 | 18934 | 0.000 |
| Difference in Heading (Deg) | .975 | 238.416 | 2 | 18934 | 0.000 |

Table 6.2 Tests of Equality of Group Means Results using different Independent Variables between Car, Train & all other Transportation Modes as a Third Category

We could comfortably conclude from this statistical evaluation that speed and acceleration are the best IVs for discriminating between different transportation modes, given the specifications of the data collected in this research. We can also conclude that each variable is better at discriminating certain categories. On the other hand, using two variables that are highly correlated will bias the inference results. Section 6.6 discusses the results of the inference model using each of these IVs by quantifying the difference in the classification accuracy for each mode.

6.3 Transportation Mode Classification

This section highlights the method-related limitations in previous attempts to infer the transportation mode. This section also describes the framework used to classify the GPS segments into transportation modes. The framework is based on a SVM classification problem based on the speed and acceleration of the trajectory, as proven to be the best IVs due to the statistical evidence discussed in section 6.2. The framework uses an innovative sliding window approach to learn and classify the data instances separately for each variable. A transition matrix is later applied to amend the sequence of consecutive trip stages. A preliminary segmentation process is applied afterwards, based on the idea that a *walk* stage mostly exists as a transition between every two other stages in any trip. This enables a further reasoning on the final classification of non-*walk* stages using a trajectory clustering technique carried out next in chapter 7.

6.3.1 Classification Limitations

The range of the methods used to infer the transportation mode from GPS data has extended from logical procedural to Machine Learning (ML) approaches in order to resolve a classification problem. Stopher et al. (2008) uses a process of elimination of different modes at different phases of the algorithm. Schüssler and Axhausen (2009) developed an open source fuzzy logic engine using the median of speed, the ninety-fifth percentile of the speed and the acceleration distributions as fuzzy variables. Several studies employ decision trees to perform this classification, either alone or integrated with other techniques, such as Hidden Markov Models (HMM) (Zheng, et al., 2008; Reddy, et al., 2010; Manzoni, et al., 2010; Stenneth, et al., 2011).

A slight limitation is that the majority of these studies only consider a limited number of transportation modes. Some use as few as 3 modes (Liao, et al., 2007; Yang, et al., 2003), while most studies exclude the train and underground modes. Others generalise the motorised modes together (Reddy, et al., 2010), grouping bus and car modes.

A common practice is to start the process by segmenting the GPS track into trips, based on either a “dwell time” period (Stopher, et al., 2008a), a threshold of time without fix. Other studies go a step further by segmenting each trip into stages, identifying the change points of mode switches. However, some of these studies start by performing stage-level segmentation and then perform the classification based on the identified stages (Schüssler & Axhausen, 2009; Zheng, et al., 2008). This exerts a shortcoming in that the classification accuracy is highly reliant on the segmentation’s efficiency. On the contrary, if a *car* stage was identified as two segments, based on the fact that it moved from a speedy main road to a highly busy street, the latter might be misclassified accordingly.

Other studies that are not dependent on segmentation classify each GPS segment individually into a transportation mode and not classifying the consequent segments as a block, i.e. the change in a trajectory's motion across several segments. Even studies that perform segmentation beforehand tend to ignore this consequence across the mode switch points.

Most studies also assume that any two stages are always separated by a *walk* stage. This, while true for most cases, might fail in cases of *cycling* or driving the *car* out of a *train* station's car park for example. A useful way to account for this is to use a transition matrix to verify the mode switch between consecutive stages according to a probability matrix of such switches (Zheng, et al., 2008).

A couple of studies also use temporal information for mode inference. Liao, Fox and Kautz (2007) use the time of day to use in a probability model building assumptions about the participant's context. While this might be a useful technique to identify different activities, it might not be applicable to participants that have abnormal working hours for example. Stenneth et al. (2011), on the other hand, depends on live bus and train times information to make some inferences too, which would require a continuous input of such information for any period of time.

The classification phase of this inference framework that we propose in this work is based on SVMs to classify GPS segments into respective transportation modes. An advantage of using SVMs over other ML methods is that they can be easily trained and are applied directly to the data without the need for a feature extraction process. This allows us to learn from the structure of the data. The proposed method uses a moving window across every group of consecutive segments in order to capture the nature of participants' movements through different transportation modes. We consider all the possible transportation modes, while testing the algorithm to avoid any mode aggregations or exclusions. A segmentation process is applied to the classified data after the initial SVM inference is performed to avoid the reliance on the segmentation accuracy if we have had applied the segmentation before the classification. We also avoid using any temporal assumptions to ensure the robustness of our algorithm over different samples. A transition matrix is also applied to assign modes in the case of potential transitions between any two non-walk stages. The rest of this section provides a detailed account of our proposed framework and a summary of the chosen SVM model, while results and some discussions are presented in sections 6.6 and 7.1.

6.3.2 Support Vector Machines (SVM) Classification and Model Selection

A Support Vector Machine (SVM) is a non-probabilistic binary linear classifier. A SVM constructs a hyperplane or a set of hyperplanes in a high- or infinite-dimensional space to achieve the largest separation between different classes (Steinwart & Christmann, 2008). SVMs use an implicit mapping of the input data into a high-dimensional feature space, defined by a kernel function (a function returning the inner product $\langle \Phi(x), \Phi(x') \rangle$ between the images of two data points x, x' in the feature space). The learning then takes place in the feature space, and the data points only appear inside dot products with other points.

The kernel functions return the inner product between two points in a suitable feature space, hence defining a notion of similarity. Kernel functions do this with little computational cost

even in very high-dimensional spaces, since it does not involve any actual computations in that high-dimensional space, which is a major advantage of using SVMs. In this research, we use a Gaussian Radial Basis Function (RBF) kernel (Equation 6.6). The Gaussian and Laplace RBF kernel is a general-purpose kernel used when there is no prior knowledge about the data.

$$\mathbf{k}(\mathbf{x}, \mathbf{x}') = \exp(-\sigma \|\mathbf{x} - \mathbf{x}'\|^2) \quad \text{Equation 6.6}$$

When classifying, support vector machines separate the different classes of data by a hyperplane contained by the decision function in Equation 6.7.

$$f(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b) \quad \text{Equation 6.7}$$

And the SVM solution \mathbf{w} has an expansion presented in Equation 6.8, in terms of a subset of training patterns that lie on the margin. These training patterns, called support vectors, carry all relevant information about the classification problem.

$$\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i) \quad \text{Equation 6.8}$$

The optimal hyper-plane (Vapnik, 1998) will be the one with the maximal margin of separation between two classes. In order to extend this binary SVM into the multi-class problem, there have been reformulations of the support vector quadratic problem that deal with more than two classes. One of these reformulations, introduced by Crammer and Singer (2002) and referred to as “spoc-svc”, works by solving a single optimization problem including the data from all classes. The algorithm is presented in Equation 6.9.

$$\text{Minimise } t(\{\mathbf{w}_n\}, \epsilon) = \frac{1}{2} \sum_{n=1}^k \|\mathbf{w}_n\|^2 + \frac{C}{m} \sum_{i=1}^m \epsilon_i \quad \text{Equation 6.9}$$

$$\text{Subject to: } \langle \Phi(\mathbf{x}_i), \mathbf{w}_{y_i} \rangle - \langle \Phi(\mathbf{x}_i), \mathbf{w}_n \rangle \geq b_i^n - \epsilon_i \quad (i = 1, \dots, m) \quad \text{Equation 6.10}$$

where the decision function is:

$$\text{argmax}_{n=1, \dots, k} \langle \Phi(\mathbf{x}_i), \mathbf{w}_n \rangle \quad \text{Equation 6.11}$$

where:

m = number of training patterns,

C = cost parameter.

The cost parameter C of the SVM formulation in Equation 6.10 controls the penalty paid by the SVM for misclassifying a training point and thus, the complexity of the prediction function. A high cost value C will force the SVM to create a complex enough prediction function to misclassify as few training points as possible, while a lower cost parameter will lead to a simpler prediction function. The best C selected was found to be of value 3, where they generated the best results. This value is not too small where it allows less error in training (due to GPS errors), and since the data is very inseparable, yet it also is not too large that the model is over fit. A k -fold cross validation on the training data of value 3 is performed to assess the quality of the model (the accuracy rate for classification).

Another advantage of SVMs and kernel functions is that the selected kernel could be applied directly to the data without the need for a feature extraction process. This is particularly important in problems where a lot of structure of the data is lost by the feature extraction

process (e.g. the sequence of a GPS trajectory's movements: such as the way a *car* can move fast, stop for traffic and then move again).

6.4 Window-Based SVM Classification

The loss of GPS coverage due to indoor activity causes the track to be filled with long gaps with no movement till the first point that follows that gap. Therefore, the first step is to segment the track due to these gaps, as an initial segmentation process based on long stays with very little movement (no spatial clustering involved). This indoor segmentation acts only as a preliminary segmentation, whereas an elaborated indoor segmentation process is performed after classification in chapter 7. Nevertheless, after this preliminary segmentation, the data is then ready and prepared for the SVM learning process and classification, being a supervised learning framework.

6.4.1 Multi-Segment Instance Classification

As previously mentioned, a SVM constructs a hyperplane in a high-dimensional space to achieve the largest separation between different classes, where the higher the dimension, the better the separation. Consequently, SVM maps original finite-dimensional space into a much higher-dimensional space to increase the separation. In this work, we enter the classification with more than one dimension in order to have a far better separation to start with.

Since we only have one dimension to begin with (speed or acceleration), we aim to simulate a multi-dimensionality to study sequences of GPS trajectory movements rather than each segment on its own (e.g. the stop and go motion of a car due to traffic). Therefore, the data is divided into equal-sized instances of several segments as demonstrated in Figure 6.2b. This simulates the multi-dimensionality of the data in the learning process which is an advantage of SVM where it eliminates the need for feature extraction, as mentioned in the previous subsection. The main reason for using instances is that it is more meaningful to study a certain stage of a trip than one single segment value; this exposes the learning process to consequent GPS data that represent the variability in one's manner when undertaking each transportation mode.

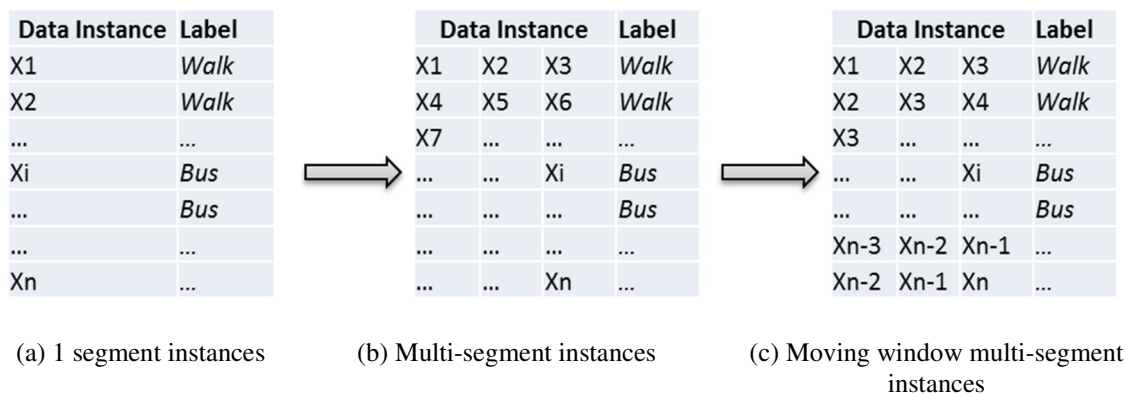


Figure 6.2 Division of Data into Equal-Sized Instances (Three in this Case)

For that purpose, the data is divided into two thirds for learning and one third for validation purposes. Data instances are then formed out of the learning data. The data instances then enter the SVMs learning process using the stationary Gaussian kernel with a radial basis kernel function (RBF) using the multi-class method.

Figure 6.3 shows several window (instance) sizes that were tested. A suitable size from 3-8 segments was identified to be the most adequate. As might be noted, the classification gives better results for longer data instances. However, a longer sequence of mixed transportation modes could introduce higher complexity, since the probability of having several modes within one instance is introduced, which will over-complicate the classification problem. Therefore, we chose to use the small-sized instance that still contains a decent number of segments to represent a realistic sequence; in this case three.

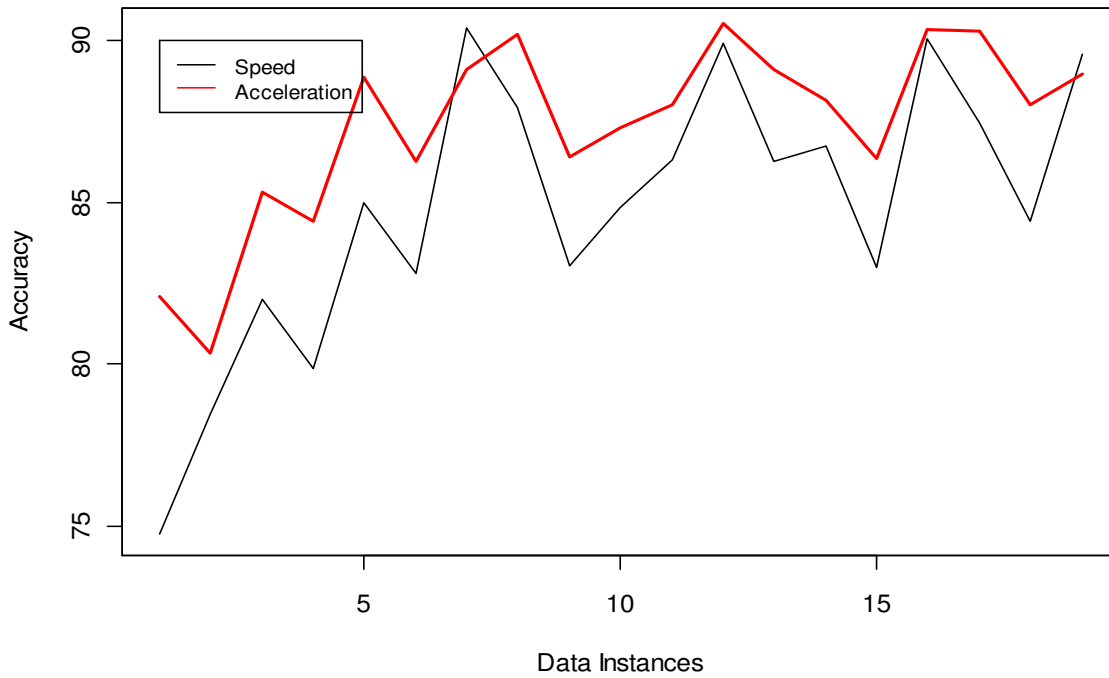


Figure 6.3 SVM Classification Accuracies using Different Lengths of Data Instances

The multi-segment instance classification achieves around an 80% inference accuracy using either speed or acceleration. This is shown in the confusion matrix in Table 6.3, where the red colour lightness varies according to accuracy of the classification for the diagonal axis (darker lightness (e.g. car) reflects higher classification accuracy than brighter lightness (e.g. bus)). The rest of lightness variance in Table 6.3 reflects the confusion in classification between different classes, with a darker lightness reflecting a higher confusion (e.g. nearly 40% of bus mode class is classified as car). There appears to be a good discrimination between the train mode and the rest, yet having a great confusion with the bus mode. The other classes seem to perform well, except the bus and tube modes, since the latter often consists of only one segment and therefore, it is merged into stages that are dominated by other modes. The classification, however, is non-realistic due to the assumption that the track is segmented into similar-mode stages.

| | | Actual | | | | | | Total Count |
|----------------|--------------|------------|------------|--------------|--------------|-------------|-------------|-------------|
| | | <i>bus</i> | <i>car</i> | <i>cycle</i> | <i>train</i> | <i>tube</i> | <i>walk</i> | |
| Classification | <i>bus</i> | 27.03% | 6.30% | 26.01% | 52.88% | 11.11% | 0.35% | 180 |
| | <i>car</i> | 39.86% | 76.72% | 11.56% | 2.88% | 55.56% | 2.12% | 523 |
| | <i>cycle</i> | 25.00% | 9.35% | 57.80% | 0.00% | 18.52% | 0.09% | 192 |
| | <i>train</i> | 0.68% | 0.38% | 0.00% | 44.23% | 0.00% | 0.00% | 49 |
| | <i>tube</i> | 4.05% | 4.96% | 0.58% | 0.00% | 3.70% | 0.27% | 37 |
| | <i>walk</i> | 3.38% | 2.29% | 4.05% | 0.00% | 11.11% | 97.17% | 1125 |
| Total Count | | 148 | 524 | 173 | 104 | 27 | 1130 | 2106 |

Table 6.3 Confusion Matrix for Classification of Instances of 3 Segments

6.4.2 Moving Window SVM Classification

In order to allow going into the segment level rather than merging different modes into the same stage, we applied a fixed-length moving window on the whole track; sliding that window segment-by-segment along the track's speed values once and once more for acceleration. Every time the window slides, a classification of that instance of data is performed. Figure 6.2c and Figure 6.4 illustrate this process, where a moving window classifies each 3-sized instance moving segment-by-segment along the track.

Once the classification is performed on the instances, the classification is passed over to the segment level and the change points in the track are identified. The change points are initially identified as any two consequent instances with different modes; the first mode being *a* and the second *b*. Then, the algorithm mines into the last instance with the mode *a* and assigns the classification of the first and second segments as *a*, and the third's as *b*. The same happens with the first instance with the *b* mode, passing the classification of the first segment as mode *a* and the second and third segments as mode *b*.

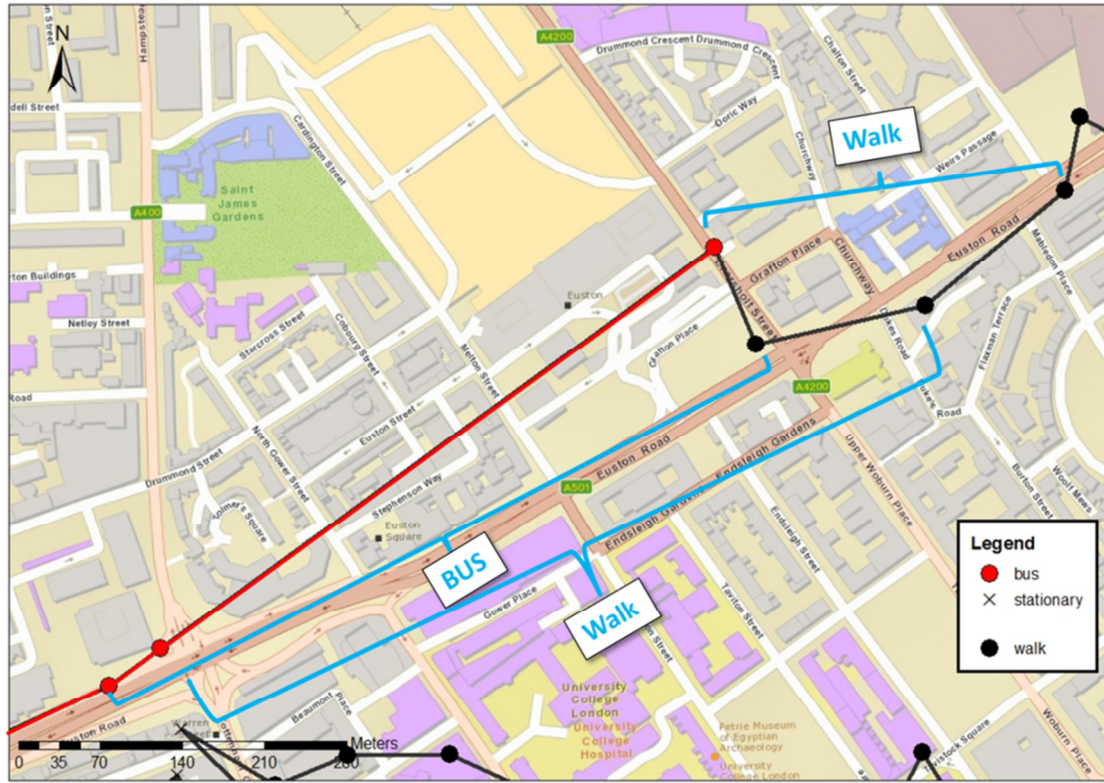


Figure 6.4 Moving Window Classifying each 3-Segment Instance moving segment-by-segment along the track

6.5 Verification by Initial Modal Segmentation

The framework then applies a verification process to each classified arc. It does this by applying two processes iteratively. The first of these two processes runs through each change point in the segment level and assesses the probability of mode *a* and *b* following each other according to a transition matrix (Table 6.4). This matrix is based on Zheng et al. (2008) and is compiled from this research's data. The matrix contains the different probabilities of switching between every two modes, which is a good indication of the natural flow of modal mixes.

| Transportation modes | Bus | Car | Cycle | Train | Tube | Walk |
|----------------------|------|------|-------|-------|------|------|
| Bus | | 0.9 | 0 | 0 | 0 | 99.1 |
| Car | 0 | | 0 | 2.2 | 0 | 100 |
| Cycle | 0 | 0 | | 1.3 | 0 | 97.8 |
| Train | 7.1 | 0 | 7.1 | | 0 | 85.7 |
| Tube | 0 | 0 | 0 | 1.5 | | 98.5 |
| Walk | 29.5 | 37.3 | 11.8 | 3.2 | 18.2 | |

Table 6.4 Transition Matrix between Modes showing Probabilities of different Modal Mixes occurring (%)

As could be noted from Table 6.4, almost all modes are followed by a *walk* mode. Therefore, the algorithm then segments the track into several stages, where every two different modal stages are separated by a *walk* stage. However, some stages will have two or more modes. In this case, the most dominant mode will be assigned to the whole stage, unless in the case of two modes, the ratio is less than 1:2 between the segments of *a* and *b* or vice versa. This creates a continuous flow of modes along different periods of the track.

As we demonstrated earlier in section 6.2.1, we use two IVs to conduct this classification; namely speed and acceleration. Therefore, we run the classification framework once for speed and once for acceleration assessing the performance of each of the variables in the process. We integrate the results of the best mode results, obtained from one variable, with the best from the other. This relies on the fact that each variable would be a better discriminator for some modes over the others. Section 6.6 describes this integration in details along with the results obtained from each variable.

6.6 Classification Results

Building on our previous work, we consider acceleration with the speed for this classification problem. The results of the moving window algorithm using speed reveal an accuracy of 72% and 83% using acceleration. This demonstrates a considerable improvement over both the previous accuracies of the multi-segment instances, without applying a moving window approach. It also has the advantage of classifying on a segment level-basis, rather than only classifying instances. Table 6.5 and Table 6.6 show the confusion matrices of this classification using speed and acceleration respectively. Some speed classification errors could be noted, such as the *car* mode with other transportation modes, while some modes, such as the *walk* mode, seem to be better classified using speed.

| | | Actual | | | | | | Total Count |
|----------------|-------|--------|--------|--------|--------|--------|--------|-------------|
| | | bus | car | cycle | train | tube | walk | |
| Classification | bus | 31.96% | 6.84% | 5.09% | 0.00% | 6.15% | 4.98% | 1348 |
| | car | 43.36% | 63.32% | 7.35% | 12.34% | 39.75% | 8.43% | 4742 |
| | cycle | 16.28% | 1.57% | 85.31% | 0.97% | 18.03% | 6.75% | 2354 |
| | train | 0.60% | 20.41% | 0.13% | 81.01% | 2.05% | 0.83% | 1815 |
| | tube | 1.88% | 4.38% | 1.10% | 5.69% | 30.33% | 0.12% | 387 |
| | walk | 5.93% | 3.48% | 1.03% | 0.00% | 3.69% | 78.90% | 8290 |
| Total Count | | 1333 | 4708 | 1552 | 932 | 244 | 10167 | 18936 |

Table 6.5 Confusion Matrix of Moving Window Algorithm based on Speed

| | | Actual | | | | | | Total Count |
|----------------|-------|--------|--------|--------|--------|--------|--------|-------------|
| | | bus | car | cycle | train | tube | walk | |
| Classification | bus | 42.31% | 0.66% | 8.12% | 0.00% | 18.85% | 2.59% | 1030 |
| | car | 18.23% | 87.26% | 10.50% | 12.77% | 18.44% | 4.77% | 5163 |
| | cycle | 21.83% | 1.10% | 76.87% | 1.93% | 11.07% | 1.91% | 1775 |
| | train | 0.15% | 5.40% | 0.06% | 84.01% | 0.00% | 0.35% | 1076 |
| | tube | 3.53% | 2.95% | 0.90% | 1.29% | 33.20% | 1.41% | 436 |
| | walk | 13.95% | 2.63% | 3.54% | 0.00% | 18.44% | 88.97% | 9456 |
| Total Count | | 1333 | 4708 | 1552 | 932 | 244 | 10167 | 18936 |

Table 6.6 Confusion Matrix of Moving Window Algorithm based on Acceleration

In order to get a better accuracy measure for the classification, we perform an inter-reliability analysis using the Kappa statistic to determine consistency among coders. Cohen's Kappa is generally thought to be a more robust measure than simple percentage agreement calculation, since K takes into account the agreement occurring by chance (Carletta, 1996). The Kappa coefficient (K) measures pairwise agreement among a set of coders making category

judgments, correcting for expected chance agreement, and hence is thought to be a good measure of any classification's accuracy. Kappa is calculated from Equation 6.12.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad \text{Equation 6.12}$$

where:

$P(A)$ = the proportion of times that the coders agree,

$P(E)$ = the proportion of times that the coders we expect them to agree by chance.

As illustrated in Table 6.7, the inter-rater reliability for speed was found to be 0.586 ($p < 0.001$), 95% CI (0.578, 0.594) and for acceleration 0.743 ($p < 0.001$), 95% CI (0.735, 0.751). That is to say, K values reflect a moderate agreement for speed and a substantial agreement for acceleration, according to rule of thumb values of Kappa (Landis & Koch, 1977).

| IV | | Value | Asymp. Std. Error ^a | Approx. T ^b | Approx. Sig. |
|--------------|----------------------------|-------|--------------------------------|------------------------|--------------|
| Speed | Measure of Agreement Kappa | 0.586 | 0.004 | 139.899 | .000 |
| | N of Valid Cases | 18936 | | | |
| Acceleration | Measure of Agreement Kappa | 0.743 | 0.004 | 167.029 | .000 |
| | N of Valid Cases | 18936 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Table 6.7 Symmetric Measurements for Cohen's Kappa Values for Speed & Acceleration

6.6.1 Type I and II Errors

Table 6.8 shows the difference between the accuracies obtained from classification using acceleration and speed classification. The red values express the excellence of acceleration classification over speed and vice-versa for the blue values. As previously noted, it seems very obvious that some modes, such as the *car* mode, are better identified using acceleration and less misclassified as other modes. On the other hand, other modes, such as *walk*, are less confused for using speed. On the other hand, the rest of the modes have little difference in results from using either IVs. An interesting confusion, however, occurs between the *train* and the *car* results, where speed appears to have a better performance by not classifying *train* as *car* (8% better), while acceleration performs better without confusing *car* instances with a *train* classification (20% better).

| | | Actual | | | | | |
|----------------|-------|---------|---------|--------|--------|---------|--------|
| | | bus | car | cycle | train | tube | walk |
| Classification | bus | 10.35% | -6.18% | 3.03% | 0.00% | 12.70% | -2.38% |
| | car | -25.13% | 23.94% | 3.16% | 0.43% | -21.31% | -3.66% |
| | cycle | 5.55% | -0.47% | -8.44% | 0.97% | -6.97% | -4.84% |
| | train | -0.45% | -15.02% | -0.06% | 3.00% | -2.05% | -0.47% |
| | tube | 1.65% | -1.41% | -0.19% | -4.40% | 2.87% | 1.29% |
| | walk | 8.03% | -0.85% | 2.51% | 0.00% | 14.75% | 10.07% |
| Total Count | | 1333 | 4708 | 1552 | 932 | 244 | 10167 |

Table 6.8 Acceleration & Speed Confusion Matrix Result Difference

It could also be noted from the main diagonal of Table 6.8 that *bus*, *car* and *walk* are better classified using acceleration, while *cycle*, *train* and *tube* are higher classified using speed. This does not mean that each group of modes should be classified using their respective IV, but it only suggests that they are over classified using these specific IVs. The trick here is to select the IV that better discriminates the classified mode from the rest. That is to be achieved by selecting the IV that achieves a higher classification for each mode, while not over-classifying that specific mode and hence, decreasing the accuracy of the other modes. This would also have the advantage of accounting for the effect of the sample size of each mode. This could be achieved by testing whether a certain variable on average dominates the row and column of each mode in Table 6.8 (actual and classified mode), while if acceleration dominates in the column level (Type I error) but speed dominates in the row level (Type II error) (such as *walk*), that would mean that acceleration is only over-classifying that specific mode. This calculation results in Table 6.9, where each mode is assessed for the Type I and II errors' excellence of one IV over the other, given that red represents an excellence of acceleration and blue for speed.

| Mode | Type I Error Difference | Type II Error Difference | Type I & II Errors Difference |
|---------|----------------------------|-----------------------------|-------------------------------------|
| bus | -10.35% | 7.16% | -3.19% |
| car | -23.94% | -46.52% | -70.45% |
| cycle | 8.44% | -5.76% | 2.68% |
| train | -3.00% | -18.05% | -21.06% |
| tube | -2.87% | -3.08% | -5.95% |
| walk | -10.07% | 24.44% | 14.37% |
| Average | -6.97% | -6.97% | -13.93% |

Blue (+ve) values demonstrate Speed's excellence
Red (-ve) values demonstrate Acceleration's excellence

Table 6.9 Acceleration & Speed Results Type I & II Error Difference

6.6.2 Integration Results

As can be noted from Table 6.9, acceleration seems to produce better results than speed for most transportation modes with the exception of *walk* and *cycle*, achieving an average supremacy of nearly 14% accuracy over speed. The inter-rater reliability for the raters was found to be $K = 0.802$ ($p < 0.001$), 95% CI (0.794, 0.810), which reflects almost perfect agreement. We adopted these results into our final integrated result of the inference, resulting in an accuracy of 88%. Table 6.10 shows the confusion matrix of this integration, demonstrating a better separation specifically for the *car*, *train* and *walk* modes.

| | | Actual | | | | | | Total |
|----------------|--------------|------------|------------|--------------|--------------|-------------|-------------|-------|
| | | <i>bus</i> | <i>car</i> | <i>cycle</i> | <i>train</i> | <i>tube</i> | <i>walk</i> | Count |
| Classification | <i>bus</i> | 58.29% | 1.02% | 11.73% | 0.00% | 19.67% | 1.54% | 1212 |
| | <i>car</i> | 15.75% | 88.47% | 9.21% | 12.77% | 10.25% | 2.57% | 4923 |
| | <i>cycle</i> | 10.58% | 1.08% | 75.19% | 1.93% | 6.15% | 1.41% | 1535 |
| | <i>train</i> | 0.15% | 4.76% | 0.06% | 84.01% | 0.00% | 0.11% | 1021 |
| | <i>tube</i> | 3.00% | 1.72% | 0.58% | 1.29% | 45.49% | 0.55% | 309 |
| | <i>walk</i> | 12.23% | 2.95% | 3.22% | 0.00% | 18.44% | 93.82% | 9936 |
| Total Count | | 1333 | 4708 | 1552 | 932 | 244 | 10167 | 18936 |

Table 6.10 Integrated Acceleration-Speed Moving Window Algorithm Confusion Matrix

Some modes appear to be performing better than others. We could note from Table 6.10 that the *car*, *train* and *walk* modes are discriminated very well using this classification. In contrast, the *cycle* mode seems to be classified moderately while the *bus* and *tube* modes still require enhancement. This could be carried out using a network matching process to both the *bus* and *tube* networks. This further work is discussed in chapter 8 for enhancing the classification of these two latter modes.

6.7 Summary

In this work we discuss the classification problem of inferring transportation mode from sparse GPS data. We first provide the means for assessing the significance of each potential independent variable that could be used for this process. We provide a statistical evaluation using the data collected by this research within Greater London as a case study. The outcome of this process provides evidence that speed and acceleration are the favourable candidates to undergo this classification problem showing a great discriminatory power in this context. However, each of these variables is also proven to be fit for identifying certain modes; *car* mode being better identified using acceleration and *walk* using speed as examples.

Building on previous attempts and on the results of the statistical evaluation, we provide in this study a novel approach for inferring the transportation mode from sparse GPS data without any extra information. In contrast to existing techniques, our approach uses one consistent framework based on support vector machines (SVM) to classify each segment into its respective transportation mode. Unlike many previous attempts, the framework tends to study the whole pattern of the trajectory motion during the whole trip using the advantage of being an offline process. The framework does this by first classifying several consequent segments together (named as an instance) with a certain window size, and sliding this window along the whole track classifying each instance. The most adequate window size was found to be of 3 segments length. The classification is then assigned to the segment level to each of the segments participating in each instance. In order to preserve the cohesiveness of the classification of the track, we segment the track into stages of different modes, each two stages separated by a *walk* stage, except for certain scenarios where we then apply a transition matrix to assess the modal mix occurrence probability.

Our model achieves relatively good accuracies using either speed or acceleration. However, building on the findings of the statistical evaluation and the SVM classification, results from the classification using both speed and acceleration are combined together. This is based on the fact that each variable is better at classifying certain modes. Finally, an accuracy of 88% is achieved from the combined result at segment level with a Kappa statistic reflecting almost perfect agreement. A good segmentation is also achieved between different modal stages; which enhances the accuracy of the classification.

Further work is required to further separate modes of similar speed/acceleration patterns, such as *bus* and *tube* modes using network matching from the rest. This is carried out and presented in chapter 8, describing a novel approach to implementing network matching in the highly complex London multi-modal network environment. Another finding is that the accuracy of the device being used and its firmware also appeared to have some effect on the classification results that could be explored in further work.

6 Phase I: Classification (Moving Window-Based Support Vector Classification (SVC))

Chapter 7

Phase II: Segmentation (Identifying Stops & Change Points)

7 PHASE II: SEGMENTATION (IDENTIFYING STOPS & CHANGE POINTS) ¹¹

This chapter describes the segmentation as the second phase constituting the inference framework defined in of Chapter 3. The first phase of classification has been described in Chapter 6, and the third phase of network matching will be described in Chapter 8.

In this research, segmentation refers to breaking a GPS track into stages and stops by respectively identifying trip mode switches and stops within stages (modes). As explained in chapter 2, mode switches and stops can be treated interchangeably in the context of detecting the mode of transport, and hence, we will often refer to them as stops and gaps. The natures of those two are different which requires us to detect each of them using different clustering approaches. The importance of detecting stops and gaps to this research lies in their high frequency within any GPS track. Moreover, another importance is due to their often occurrence as an intermediate transition between other mode types within a typical GPS track.

As mentioned in chapter 2, ignoring such **stops** has been a limitation in previous studies attempting to detect the mode of transport from GPS data. Ignoring stops leads to the increase of errors due to the introduction of slow speed values in other modes during classification. Another problem is the frequent confusion between the walk and the stationary modes. In section 7.1, we attempt to identify stops within a GPS data track using a **spatio-temporal clustering** method to identify stops (or stationary mode segments). This method overcomes previous clustering techniques by adding a temporal element as opposed by previous clustering techniques used that were solely spatial in nature. The results reveal an accuracy of around 90% for identifying stops, which is increased when additional reasoning is further applied.

Chapter 2 also highlighted attempts to identify **static indoor activity** or **gaps**. The spatio-temporal clustering algorithm we develop for identifying stops also identifies indoor activity in situations of partial GPS signal blockage. However, due to GPS cold start problems, a challenging issue is identifying the exact location of these activities in situations of total GPS signal blockage. Solutions such as to increase the data collection rate highly consumes the device battery, device memory, and computation cost. However, using multi-day data, a possibility emerges to infer the details of a bad-data trip by imputing the details from the same trip made on another day of the week. Consequently, we develop a purely **spatial clustering** algorithm based on k-means classification followed by identifying the centroids of the identified clusters using the mean of all GPS points within each cluster. The second challenging issue with static indoor activity is identifying its occurrence in the first place. For that purpose, we adopt the usage of the concept of using a dwell time (1-5mins) to identify static indoor activity in situations with total GPS signal blockage. We also assess the accuracy that this approach achieves by comparing these detected instances with the set of modal change points within the pilot data, where indoor occurrences are also counted as a change in mode of transport.

¹¹ Part of this chapter is based on a publication of ours: Bolbol, A. Cheng, T. Tsapakis, I. and Skarlatidou, A., 2011. Identifying Intermediary Modes from GPS Data. Presented at *Association of American Geographers Annual Meeting*, Seattle, Washington, April 12-16, 2011.

As a final step, we integrate the set of stops and indoor activity gaps with the SVM classification results achieved from chapter 6 to enhance the accuracy of the instance where change points occur. The accuracy of this comparison is then assessed using the pilot dataset collected in this research.

Another prolonged argument was discussed in chapter 2 regarding what is to be considered a stop at the labelling stage will lead to accurately assessing the efficiency and accuracy of identifying these stops and ends when validating the attained results. Chapter 5 has standardised the rules by which these arguments are settled. A list of significant place types was defined as well as situations where stops occur, providing a clear differentiating between both classifications when participants tag their tracks. This chapter uses data tagged by the participants according the standards set by this research, and therefore, the validation process is considered to be highly accurate.

7.1 Detecting Stops

A stop, or stationary mode, occurs when a trajectory is still and does not move. Chapter 5 described situations where a stop is decided upon to occur in this research such as a sit in the park or waiting for the bus for example. However, using a GPS sensor would usually produce an erroneous location every time a fix is attained due to systematic errors. This error could also partially be due to the brief movement of participants while still being in stationary situations. Hence, a stationary mode is not actually totally stationary, where a participant at rest could continue to move within several metres yet still be considered stationary. Moreover, GPS errors could vary from 20 to 200 meters as will be illustrated later in this subsection.

As described in Chapter 2, a variety of approaches have been developed to cluster GPS fixes in order to identify stops in a GPS track. Although most of the developed algorithms are simple, require low cost computation, and efficiently identify places where participants visit, yet they do not account for slow speed trajectories such as walks, nor account for complex transport networks. Moreover, almost all methods lack labelled data to validate the developed algorithms. Another major limitation GPS clustering methods possess is applying purely spatial clustering while ignoring the spatiotemporal nature of these trajectories, which in turn leads to missing out on shorter stops. This is mainly because most previous studies aim at identifying significant places to the participant rather than stops.

In contrast to previous research, we attempt to classify a stationary status as a type of travel mode on its own respect, as it could be an activity which forms part of the “travelling” activity (e.g. waiting for the bus). For this purpose, we apply a spatio-temporal approach to detect stops within a GPS trajectory track rather than identifying repeated daily patterns of committed trip ends. Once all stop clusters are identified, we could then assess if they are a significant stop and hence a modal change is probable or if they are an insignificant stop within a given stage.

7.1.1 Statistic Evaluation

In order to understand a stationary cluster; we need to first look into the nature of stationary data. Figure 7.1 shows a box plot for speed and distance values of stationary segments within the training dataset. The plot illustrates the distribution of GPS errors among the dataset

population. It could be noted that most of the error in positioning is between 0 and 1.5 m/s for speed and 0 and 100 m for distance. This is important to inform the clustering algorithm on the speed and distance thresholds where stops occur.

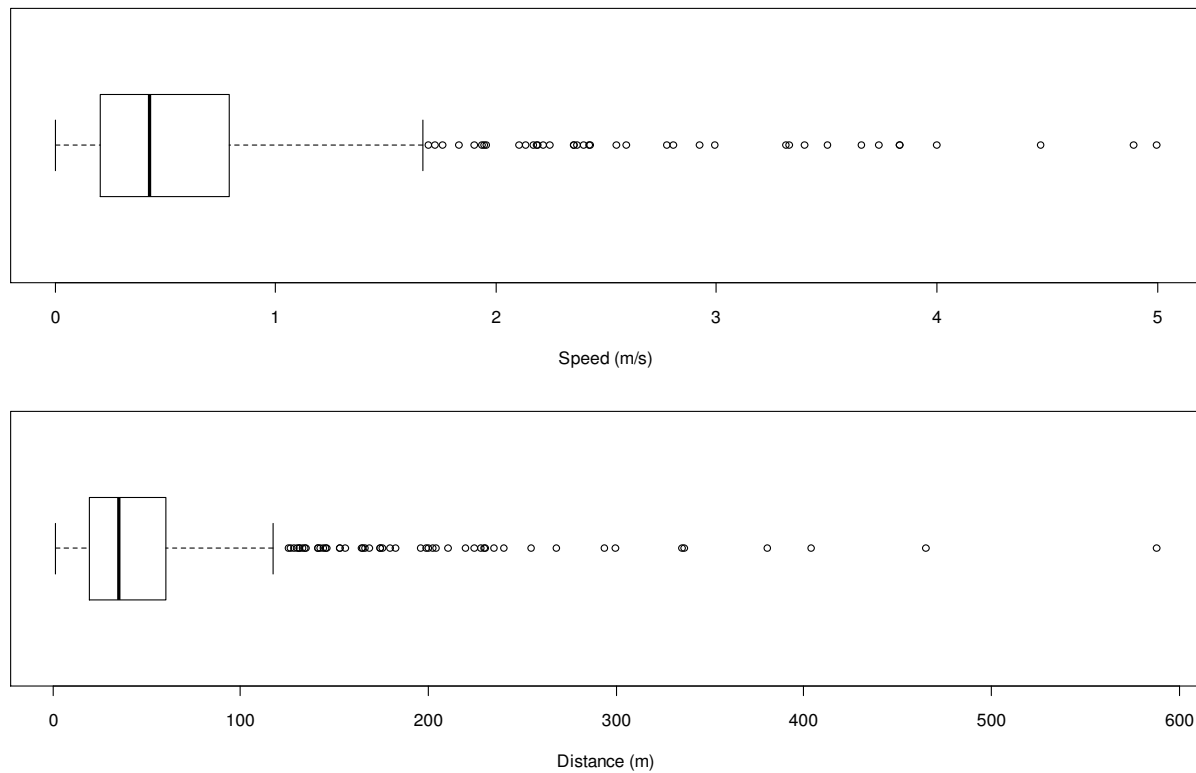


Figure 7.1 Box Plots for Speed (above) & Distance (below) Values

Table 7.1 and Table 7.2 show the mean, percentiles, inter-percentile range IQR, variance, standard deviation and specific percentile calculations stationary segments in the training dataset. As might be noted, the 90th percentile of the speed was found to be around 1.37m/s and for distance 100.19m.

| Statistical Measure | Value | | | | |
|----------------------|-------|-------|-------|-------|-------|
| Mean | 0.639 | | | | |
| Percentiles | 0% | 25% | 50% | 75% | 100% |
| | 0 | 0.215 | 0.432 | 0.794 | 4.996 |
| Range IQR | 0.579 | | | | |
| Variance | 0.481 | | | | |
| Standard Deviation | 0.694 | | | | |
| Near End Percentiles | | 80% | 90% | 98% | |
| | | 0.948 | 1.366 | 2.969 | |

Table 7.1 Statistical Calculations for Speeds of Stationary Segments

| Statistical Measure | Value | | | | |
|----------------------|----------|---------|--------|--------|---------|
| Mean | 51.955 | | | | |
| Percentiles | 0% | 25% | 50% | 75% | 100% |
| | 0.94 | 19.183 | 35.193 | 61.007 | 587.979 |
| Range IQR | 41.824 | | | | |
| Variance | 3523.608 | | | | |
| Standard Deviation | 59.359 | | | | |
| Near End Percentiles | 80% | 90% | 98% | | |
| | 71.023 | 100.188 | 229.96 | | |

Table 7.2 Statistical Calculations for Distance of Stationary Segments

7.1.2 Spatio-Temporal Clustering Algorithm

Building on this statistical evaluation, we develop an algorithm that identifies stationary clusters by constraining the cluster to speed and distance thresholds while being maintained by a number of fixes. The algorithm is written using R Project for Statistical Computation. The properties of a typical qualified stationary cluster of points identified by the algorithm include the following.

1. Speed between every 2 points should be less than a certain threshold (we use the 90th percentile of speed)
2. Distance between every 2 points should be less than a certain threshold (we use the 90th percentile of distance)
3. Two dense clusters that are separated by two segments will probably reflect an outlier GPS error, and hence, will be merged into one big cluster
4. A cluster occurs when two segments (three points) or more have been qualified as stationary
5. A cluster has a centroid that defines the most probable position of the trajectory at that cluster instance

Based on the rules set above, the algorithm we develop applies a spatio-temporal filter that identifies the described cluster situation by these rules. The algorithm goes through speed and distance data between every 2 consecutive fixes and performs a series of checks. The algorithm first moves a window through every segment from the start of the track and stops whenever a speed less than the specified speed threshold is detected. The algorithm then creates a centroid between the start and end points of that segment, which is considered to be the centre of the preliminary identified cluster. The window then moves onto the following segment and if the speed is more than the speed threshold, the previously identified cluster is reverted and the window moves on to the next segment. On the other hand, if the speed were to be still less than the defined threshold, the distance between the last point added (second point in the newly added segment) and the previous centroid is calculated. If this distance is less than the identified threshold, the cluster is declared as a stop and a new centroid is identified as the average location between all the points in the cluster. The window then moves onto the following segment and a similar reasoning is implemented. At the end, the window leaves this group of points as a stop cluster with a final centroid representing the most probable location of the stop as the average location of all points involved in the cluster.

This is illustrated in Figure 7.2 graphically as hypothetical example of a qualified stationary cluster.

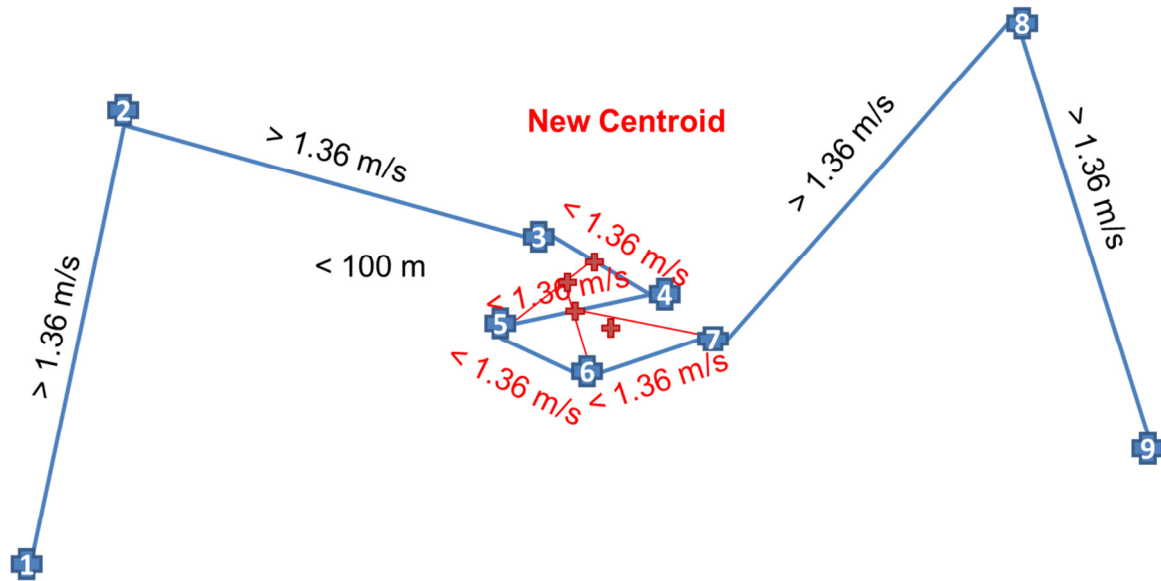


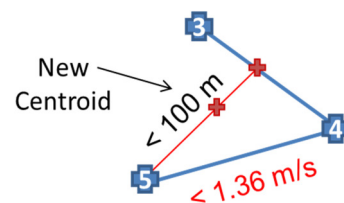
Figure 7.2 General Description of Clustering Algorithm

After all clusters in the track have been identified, a second window moves along the track searching if any two consecutive clusters are separated by two segments (three points). If the distance between the centroids of the two detected consecutive clusters is less than the threshold distance, the two clusters are merged into one. A new centroid is then calculated as an average position between both centroids. Figure 7.3 provides a detailed graphical description of the algorithm showing its step-by-step structure.

Step 1: Checks if the speed is less than a certain threshold (90th percentile of speed values) - ***Creates a centroid and a stationary cluster is created***



Step 2: Checks the following arc's speed. If less than the threshold, distance between the end point and the previous centroid is checked if it were less than a distance threshold (90th percentile of distance values) - ***Adds the new arc to the cluster, and moves the centroid to the new centre***



Step 3: Repeats the same steps till the speed rises above the speed threshold again and only then; it exits

Step 4: After all the track has been covered, it searches for any two clusters that are classified as stationary and are separated by one or two non-stationary arc, having each of these two clusters consisting of more than 1 arc. And If so, it checks if the distance between the centroids of both clusters is less than the distance threshold - ***Both the clusters are merged into one and a new centroid is created***

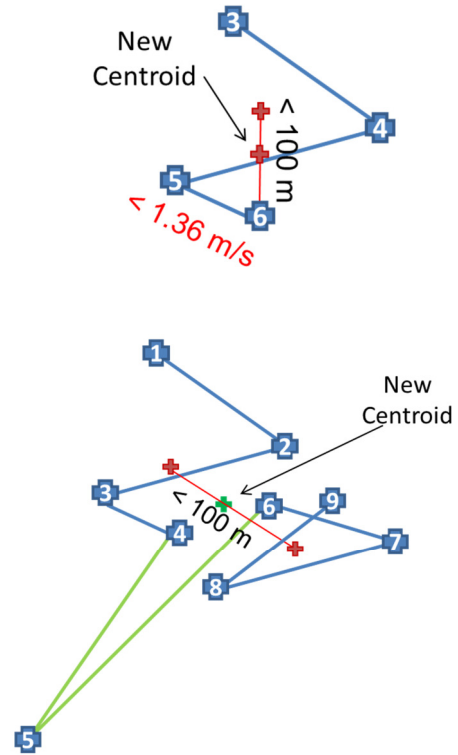


Figure 7.3 Description of the Developed Algorithm used in Clustering for Stops

7.1.3 Validation & Results

We have obtained the learning figures from two thirds of the pilot dataset while the clustering algorithm was tested on the validation third of the pilot dataset. The two thirds used for learning are then used for cross-validation to generate multiple training/validation set pairs (Alpaydin, 2004). The results obtained from the validation process reveal different accuracies as a result of using different thresholds. As a result; we establish an instrument to measure the performance based on different threshold values used by the algorithm. Two measures were identified to quantify the performance of the algorithm; namely, the clustering accuracy and the clustering error. The following equation points out that the calculation of the clustering accuracy is a measure of how well the algorithm detects stops.

$$\text{Clustering Accuracy} = \frac{\text{Stationary segments correctly detected}}{\text{All stationary segments}} \quad \text{Equation 7.1}$$

On the other hand, the following equation calculates the clustering error as being a measure of over-classification of stops.

$$\text{Clustering Error} = \frac{\text{NonStationary segments incorrectly detected as stationary}}{\text{All segments detected as stationary}} \quad \text{Equation 7.2}$$

The measures of accuracy and error were calculated for several speed and distance thresholds set for the clustering algorithm. Table 7.3 shows using speeds of (0.8, 1) m/s and distances of (70, 100) m as thresholds to be used in the clustering algorithm. We also compile a confusion matrix for results from different threshold values used by the clustering algorithm. Each row of the matrix represents the instances in the predicted class, while each column represents the

instances in the actual class. Table 7.4 shows a group of these confusion matrices compiled as a result of using different speed and distance thresholds used in the clustering algorithm.

| Accuracy (%) | | | Distance to Centroid (m) | | |
|--------------|-------|-------|--------------------------|-------|--|
| | | | < 70 | < 100 | |
| Speed (m/s) | < 0.8 | 85.97 | 87.58 | | |
| | < 1.0 | 87.58 | 89.4 | | |

| Error (%) | | | Distance to Centroid (m) | | |
|-------------|-------|-------|--------------------------|-------|--|
| | | | < 70 | < 100 | |
| Speed (m/s) | < 0.8 | 13.89 | 16.12 | | |
| | < 1.0 | 15.92 | 18.49 | | |

Table 7.3 Accuracy & Error Measures for using Different Speed & Distance Values

As it might be noted from both tables, the higher the speed and distance allowance the better the accuracy measure, however, the error measure increases as well concurrently. On the other hand, to reduce the error we need to reduce the allowances, and then we would also be reducing the accuracy measure. However by looking at the confusion index, most of the error occurs due to classifying stationary clusters within walk segments. That is acceptable because walk often comprises of stationary instances. Therefore, the error measure could be assigned a lower weight of importance compared to that of accuracy.

| Centroid < 70m | | | | | Centroid < 100m | | | | |
|-----------------|----------------|--------|--------|-------------|-----------------|--------|--------|-------------|-------------|
| Speed < 0.8 m/s | Classification | Actual | | | Classification | Actual | | | Total Count |
| | | NA | Stat | Total Count | | NA | Stat | Total Count | |
| | | bus | 92.85% | 7.15% | 1216 | bus | 91.45% | 8.55% | 1216 |
| | | car | 93.85% | 6.15% | 2210 | car | 92.31% | 7.69% | 2210 |
| | | cycle | 94.98% | 5.02% | 1455 | cycle | 94.02% | 5.98% | 1455 |
| | | train | 97.52% | 2.48% | 282 | train | 97.16% | 2.84% | 282 |
| | | tube | 97.81% | 2.19% | 319 | tube | 97.18% | 2.82% | 319 |
| | | walk | 68.51% | 31.49% | 2483 | walk | 61.90% | 38.10% | 2483 |
| | | stat | 14.03% | 85.97% | 7874 | stat | 12.42% | 87.58% | 7874 |
| | Total Count | 7978 | 7861 | 15839 | Total Count | 7619 | 8220 | 15839 | Total Count |

| Speed < 1 m/s | Classification | Actual | | | Classification | Actual | | | Total Count |
|---------------|----------------|--------|--------|-------------|----------------|--------|--------|-------------|-------------|
| | | NA | Stat | Total Count | | NA | Stat | Total Count | |
| | | bus | 91.45% | 8.55% | 1216 | bus | 89.47% | 10.53% | 1216 |
| | | car | 92.81% | 7.19% | 2210 | car | 90.95% | 9.05% | 2210 |
| | | cycle | 94.64% | 5.36% | 1455 | cycle | 93.33% | 6.67% | 1455 |
| | | train | 97.52% | 2.48% | 282 | train | 97.16% | 2.84% | 282 |
| | | tube | 97.49% | 2.51% | 319 | tube | 96.55% | 3.45% | 319 |
| | | walk | 61.74% | 38.26% | 2483 | walk | 53.60% | 46.40% | 2483 |
| | | stat | 12.42% | 87.58% | 7874 | stat | 10.60% | 89.40% | 7874 |
| | Total Count | 7637 | 8202 | 15839 | Total Count | 7204 | 8635 | 15839 | Total Count |

Table 7.4 Clustering Algorithm Performance Results using different Speed & Distance Threshold Values

Figure 7.4 illustrates the effect of using different stationary percentile thresholds for speed and distance calculated from the training data on the achieved accuracy and errors in identifying

stationary clusters. The choice of suitable threshold values will depend largely on the purpose which the outcome of the algorithm is going to be used. Therefore, suitable percentile values will be chosen based on whether achieving a higher accuracy or a lower error is more important to the developed framework as a whole.

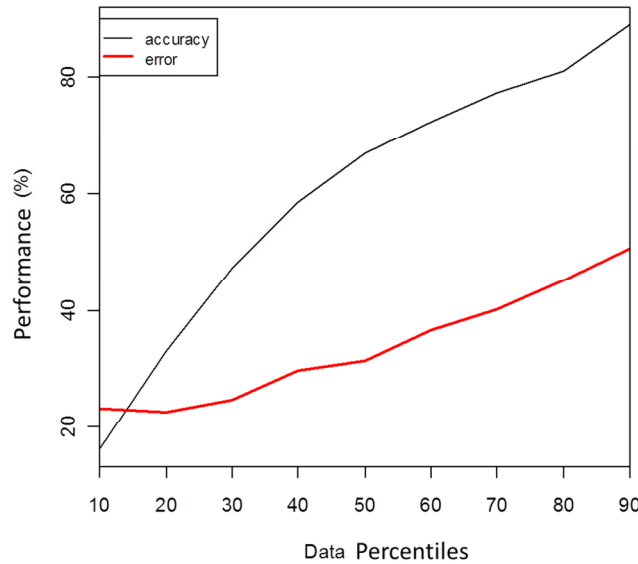


Figure 7.4 Effect of using different Percentiles of Speed & Distance Thresholds on achieved Accuracy & Error

In this framework, achieving a higher accuracy is generally of a greater importance than minimising the error level, since a further step will be added to filter out stops within a walk stage, hence, minimising the error level. As previously illustrated in Table 7.3 and Table 7.4, using values such as 1m/s for speed and 100m for distance allowances would give a fair trade-off between the accuracy and error measures for this study maximizing the accuracy (89.4%) and minimizing the error (18.49%). That is considered to be one relatively good result from an algorithmic performance viewpoint. However, when coming to perform the clustering without training data, similar values could be initially used, then new equivalent percentile values can be used every time new data is added to the training database.

7.1.4 Discussion of Results

As noted above in the previous subsection, using speed and distance thresholds that achieve higher clustering accuracy and just an acceptable level of error is beneficial enough for the inference framework as a whole. Assigning a higher significance to the accuracy measure stems from the fact that many of the false positives are actual stops that were committed during walk stages, due to the stop-and-go nature of walking. Although these clustering errors are not actual errors yet they are still considered errors due to the labelling strategy that dictates labelling the whole walk stage as solely walk mode. Therefore, these brief stops in the middle of walks are merged into the bigger walk stage as illustrated in Figure 7.5. Moreover, removing outliers from the middle of stop clusters, performed at the end of the clustering algorithm, has proven to increase the clustering accuracy and provides a better understanding of these clusters despite of GPS errors.

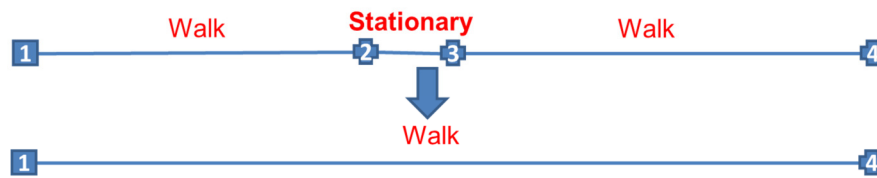


Figure 7.5 Merging Typical Stop Cluster occurrences within a walk into one Segment

7.2 Identifying Static Indoor Activity (Gaps)

In chapter 2, we have also mentioned that previous studies that attempted to identify static indoor activity. We have also highlighted that some studies attempt to identify the **exact location of indoor activity**, which is a challenge due to the GPS cold start problem. One proposed solution was to increase the data collection rate which has the limitation of highly consuming the device battery, device memory, and computation cost. Another solution was to infer the details of a bad-data trip by imputing the details from the same trip made on another day of the week for multi-day surveys, where a cold start problem did not occur. Another problem that previous research has tackled is identifying the **occurrence of the indoor activity**. Another common practice is to base the search on a threshold “dwell time” period. The value chosen for the dwell time however, will largely depend on the GPS rate of collection. Research studies were divided upon choosing a small value (1-5min) which was thought to be a “safer” choice, while others used the time of day and/or other temporal resolutions to identify the most probable destinations a participant would end their trip at.

Therefore, this section discusses our approach to infer the exact location where an indoor situation has occurred. First, we illustrate how the spatio-temporal clustering algorithm developed and applied in the previous section addresses the issue of detecting the exact location of the occurrence of indoor activity in situations of GPS signal partial blockage. On the other hand, we show how the location of indoor activity is identified in situations of total blockage using a k-means based spatial cluster algorithm that we develop in this section. Afterwards, we demonstrate how we also identify the occurrence of indoor activity in the first place in instances of total blockages when no GPS signal is detected.

7.2.1 Exact Location of Indoor Activity

Identifying the exact location where a participant in a GPS-based travel survey has entered can be difficult in situations where a large rate of GPS data is used for collection. Moreover, upon exit of indoor environments where no GPS signal is attained, a GPS cold start problem might evolve as a result of the GPS initialisation process. Figure 7.6 illustrates an indoor situation within a GPS track illustrating an example of a hypothetical track. The figure shows a track with two stages (cycle and walk) with a gap of 15 minutes between the transition point and the following point. According to the stated rule, a stop is created at the intersection of the two original stages. The stop could be a stop or a trip end as it might reflect indoor activity. The scope of this research is not to differentiate between trip ends and ordinary stops using the process of inference, rather the scope is to infer information that would aid the detection of the mode of transport. Identifying indoor activity, regardless of being a stop or trip end, would help differentiate between trip elements such as underground travel and modal switches. That is especially true in the case of detecting the occurrence of indoor activity rather than identifying the exact location of it. However, we attempt to carry out this task to identify any extra trip information that might prove achievable in the mode of transport inference process.

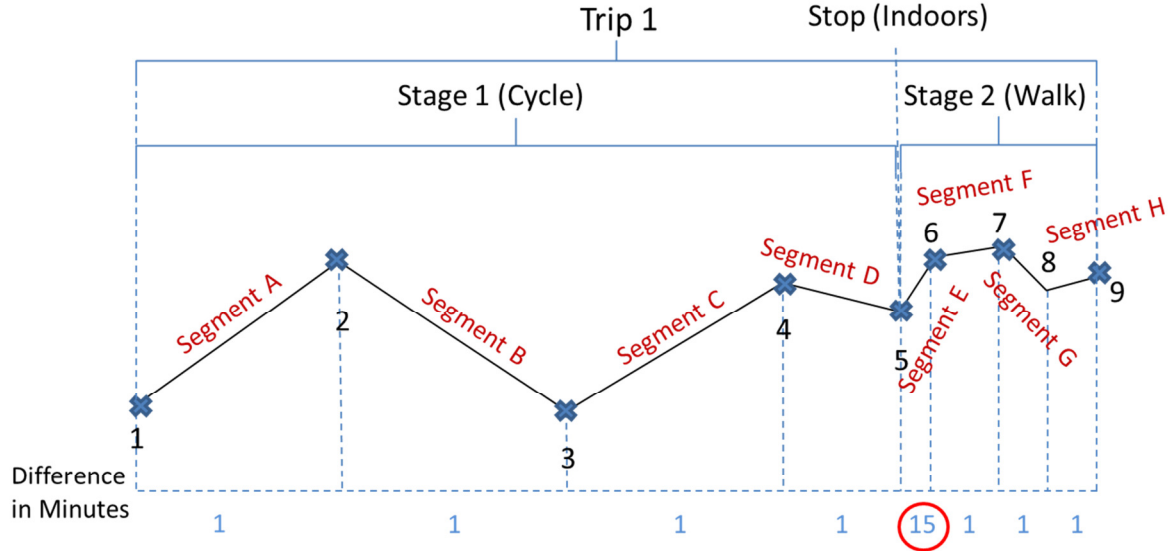


Figure 7.6 Example of Lack of Fixes between two Stages leading to creation of a Stop

As mentioned in chapter 2, an indoor stop (as we will bluntly call it), such as the one in Figure 7.6, could be in a slightly GPS-permissible environment (**partial blockage**), and hence still get position, or could **totally block** the GPS signal out (Wood & Mace, 2001). As mentioned in chapter 4, this research develops an inference framework to deal with sparse GPS data which is beneficial for minimising battery and memory consumption as well as decreasing the computational cost significantly which can enable real-time usage as a consequence. A limitation of using sparse GPS data though is decreasing the accuracy of identifying indoor locations. This algorithm developed in this research deals with partial and total blockages within indoor environments differently.

Obtaining a **partial blockage** in an indoor environment means that the spatio-temporal clustering algorithm previously developed in section 7.1 can detect the average position of the participant. An important fact to note is that the accuracy of the calculated position will most probably be directly proportional to the amount of time spent at that location, depending on the number of satellites observed. No statistical correlation testing was carried out to aid this statement since using the postcodes of significant places reported by the participants to assess a subset of their indoor destinations would hardly yield good form of assessment. This is based on the fact that a single small UK postcode may contain up to 100 addresses which can reach hundreds of metres in radius, and hence is by far is inferior to the accuracy of GPS solutions (UK Census, 2011). The results of the spatio-temporal clustering are previously reported in section 7.1 where the clustering algorithm was described and tested.

On the other hand, the spatio-temporal clustering technique cannot be applied for the **total blockage** because a total blockage means simply there is no consequent data to cluster. Therefore, once an indoor activity is detected (as described later in subsection 7.2.2); repeated daily occurrences are spatially clustered using k-means technique. K-means clustering differs from other partitioning methods by using the centroid (or mean) of the objects in the cluster as the cluster centre (MacQueen, 1967). The objective criterion used in the k-means algorithm is typically the squared-error function which is defined below in Equation 7.1 as E , where x is

the point in space representing the given GPS point, and m_i is the mean of cluster C_i . The k-means assigns each GPS point to its nearest centre forming a new set of clusters. New centres are then calculated from the newly formed clusters from GPS points in each cluster. This is replaced until the criterion function E does not change after several iterations. On the other hand, other clustering techniques such as k-medoids for example calculates the centres as the most central GPS point (Kaufman & Rousseeuw, 1990), or the Expectation Maximisation algorithm which assigns points to a cluster according to a of membership probability (Dempster, et al., 1977). K-means was chosen because the most central GPS fix would still contain GPS error, and similarly a probability would be highly biased in cases of few points due to systematic GPS errors.

$$E = \sum_{i=1}^k \sum_{x \in C_i} |x - m_i|^2 \quad \text{Equation 7.1}$$

The clustering algorithm developed in this section initially uses three clusters for each participant to enter into the k-means process. Once GPS fixes are assigned, this number is increased followed by the same k-means algorithm again. The new E value is calculated for the new classification, and this process goes on till 10 clusters. The E values are then compared, and the number of clusters with the lowest E value is chosen. The centroid of each cluster is then calculated from the GPS points in each cluster. These centroids then replace the coordinates of these gaps in the GPS track.

As previously mentioned, there is no efficient way to assess the performance of this method since postcode accuracy would be nearly equivalent to that of GPS, however, inferring the exact location of significant places is not within the exact scope of this thesis, and nonetheless its performance. Figure 7.7 shows an example of outcomes of the spatial clustering algorithm highlighting clustered locations from repeated data fixes identified over several days.

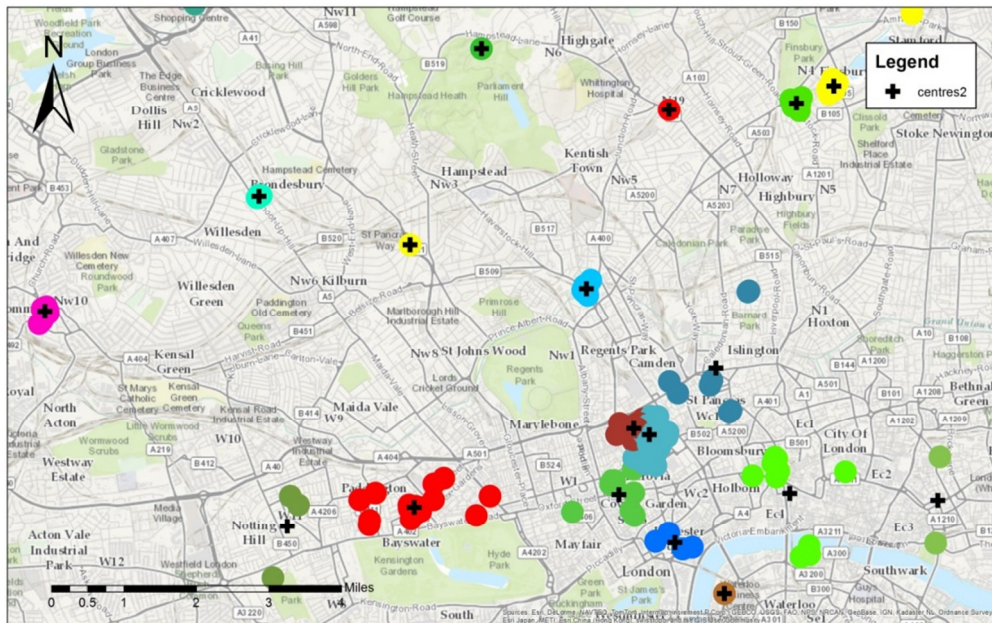


Figure 7.7 Example of Results of k-means Spatial Clustering Technique used to identify the Location of Indoor Activity Occurrences

On the other hand, an algorithm dedicated to detecting tunnel underground travel is developed in chapter 8 and applied as part of the network matching phase. The algorithm will reason about using underground travel by using a combination of a distance threshold, a time threshold and the underground network dataset. This part of network matching has the potential of enhancing the accuracy of areas of underground travel total blockage by snapping to the underground network. This also filters out the network travel from the outcomes of the rest of signal blockage instances in a GPS track.

7.2.2 Occurrence of Indoor Activity

As mentioned in the previous subsection (7.2.1), cases of indoor partial blockage can be detected using the spatio-temporal clustering algorithm we developed in section 7.1. To qualify as a stop, the algorithm decides on at least three GPS fixes. Since the GPS data is collected at 1min collection rate, as set in chapter 4, then any stop would bear a minimum of 3min wait. On the other hand, total blockage is not detected using the clustering algorithm we have developed because the stop would have happened for only 2 consecutive fixes. Therefore, a reasoning framework is applied as a second segmentation stage in order to find gaps in time series to identify total blockage occurrences. The aim of this test is to segment a participant's track according to loss of coverage, which could be interpreted as either being indoors, lost coverage for some time, or travelled using the underground tube transport. As we mentioned in the previous subsection (7.2.1), underground travel is detected in the following phase in chapter 8 using a network matching algorithm. On the other hand, this reasoning algorithm aims to specifically detect stops with no GPS coverage, or as we refer to it in this research; indoor activity. For this method we adopt the dwell time approach by using a time threshold by which we decide on whether an indoor occurrence is detected.

We use a **dwell time** of 300sec (5min) similar to previous studies, and of which is 5 times the collection rate, and hence, eliminating the chance of losing signal coverage by one fifth. We also assign a distance threshold between the two fixes as a maximum of 200m, which is 85 times the highest error level expected by GTrek devices (GTrek, 2012) used for the validation dataset of 95 participants, or 40 times the u-blox devices (u-blox, 2009) used for the pilot dataset when in open skies. Yet, this algorithm deals with GPS fixes with minimal satellite coverage, and therefore, the accuracy would easily reach these levels.

We assess this simple approach by comparing the achieved results to trip stops/ends previously identified by the participants. This very simple algorithm achieves a remarkable accuracy of nearly 95% for correctly identifying beginning of new stages, which might also be referred to as modal change points. The algorithm also identified around 42% of all change points within the testing dataset. Table 7.4 shows the confusion matrix of the achieved results for the entire dataset. As can be noted from Table 7.4, this gap detection step yields a very good performance in identifying change points (78 out of 186), bearing in mind that not all change points are indoor activities. As will be described in the next section, switch points can be indoor activity, stops, or just simply a switch in mode.

| Inferred | Truth | |
|-------------------|--------------|------------------|
| | Change Point | Non-Change Point |
| Change Point | 78 | 51 |
| Non- Change Point | 108 | 2486 |

Figure 7.8 Indoor Segmentation Test Confusion Matrix Results

7.2.3 Indoor Activity Summary

This section has discussed our approach to calculate the exact location of static indoor activity in a GPS track. We illustrate how location of static indoor activity is identified in situations of total blockage using a k-means based spatial cluster algorithm that we developed in this section. This method helps overcome GPS typical problems such as cold starts. This approach can also be applied to GPS devices that are movement-activated where when not in motion they hibernate to save battery and memory clustering several time gaps in data. The section also demonstrates the usage of the dwell time approach to identify the occurrence of indoor activity in the first place in situations of total blockages achieving a supreme accuracy of around 95%.

7.3 Identifying Change Points (Integrating Classification and Segmentation Results)

As described in chapter 2, some of the research studies that aim to detect the mode of transport from GPS data search for **change points** where the mode changes, breaking the trip into several stages. Problems in these studies mainly include being based heavily on this segmentation process which might be risky since the accuracy of segmentation might not be very high. Other limitations include using sudden changes in speed and making assumptions that every two stages are separated by a walk stage which might not always be true.

Therefore in this research, we do not base any of the inference phases on the other. Alternatively, as described in the previous two sections, we use the spatio-temporal clustering and spatial clustering techniques to identify stops, indoor activity, and modal change points, and use the classification results only to identify the remaining change points in the GPS track. In section 7.1, we attempted to identify stops of which some might be change points. Moreover, in section 2.5.3, we detected indoor activity from gaps in the GPS track, which are also considered to be change points. Yet, we have still not checked the coincidence of the identified stops/gaps with the change points that might be a switch from one mode to the other. Therefore, we develop an additional step that reasons about the results from the classification stage in conjunction with the previously identified stops/gaps, and hence, enhancing the accuracy of the classification.

The algorithm first attempts to identify consecutive segments that were classified differently in the classification phase. The code produces a seamless harmonised sequence of classified segments by checking if any segment of time t and mode x is surrounded by segments $t-1$ and $t+1$ and both having a mode of y , and if found, segment t 's mode is switched from x to y . Further harmonisation is applied at the very end of the inference framework after producing the network matching results presented in chapter 8. On the other hand, after producing an initially harmonised sequence of modes in this phase, we identify the first segments of every stage. The first segment of every stage would ideally contain the first change point between

different modes. Hence, we apply a validation process based on logical reasoning, where the stops and gaps identified from the previous sections are revised in relation to these first points and checked whether they coincide at the same fix. Figure 7.9 shows an example of this comparison, illustrating an instance where there is a coincidence between SVM results and stops/gaps identified in this chapter and another instance with no coincidence. This step searches whether any non-coincidence is shifted by not more one GPS fix, and shifts the SVM classification result by one segment to conform to the identified stop/gap. The reason why we give precedence to the stop/gap results is due to the fact that gaps/stops are more likely to be the true change points in a GPS track than the results from the classification.

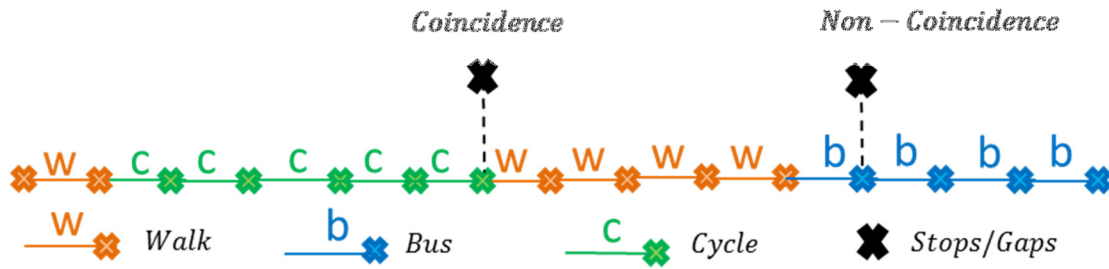


Figure 7.9 Coincidence of identified Stops/Gaps with SVM Classification Results

Results from this section achieve an accuracy of correct identification of 13 out of the 16 change points that satisfy the circumstance of a stop/gap being shifted by one fix. Although being minimal in its contribution to the overall classification accuracy, yet this result shows that such a simple step can enhance the accuracy of identifying modal change points within a GPS track significantly.

7.4 Segmentation Conclusions

This chapter discussed segmentation of a GPS track as the second phase of our approach in this thesis to infer the transportation mode from sparse GPS data. After the first phase (classification) discussed in chapter 6, the framework attempts to segment the GPS track into stops, static indoor activity, and change points. This is mainly achieved by detecting stops and gaps. Detecting stops and gaps is important for this research due to their high frequency in GPS tracks, and acting as transitional phase between other mode types within a typical GPS track. The nature of each of them is different which requires us to detect each of them using different clustering approaches. Therefore, we develop two clustering algorithms, one to deal with cases of stops or partial blockage and another with cases of total blockage (or static indoor activity).

First, we develop a spatio-temporal clustering method that applies a logical reasoning framework using a combination of travel distance and speed thresholds to identify stops within a GPS track. The algorithm also accounts for outlier GPS fixes that might occur within stop clusters. Results reveal an accuracy of around 90% for identifying stops. Identifying more clusters is considered to be more desirable than fewer clusters, where many stops occur during walking stages and are not significant to identify/differentiate in this research. For this reason, these stops in the middle of walks are later merged into the entire walk stage to eliminate confusion. Furthermore, removing outliers from the middle of stop clusters has shown to generally increase the clustering accuracy.

Static indoor activity in situations of partial GPS signal blockage could also be detected and located using the same spatio-temporal clustering algorithm. Conversely, in situations of total blockage, we use a k-means based spatial clustering algorithm to detect the location of static indoor activity. Then, we identify the centroids of the identified clusters using the mean of all GPS points within each cluster. This method overcomes typical problems of GPS technology such as cold starts. The applicability of this approach could be extended to be used with movement-activated GPS devices by clustering the resulting time gaps in data. We also use the dwell time approach to identify the occurrence of indoor activity in the first place in situations of total blockages. We assess the accuracy of this approach by comparing these detected instances with the set of modal change points within the pilot data, where indoor occurrences are also counted as a change in mode of transport. The approach reveals around 95% accuracy for detecting mode changes.

Finally, we compare the stops and static indoor activity gaps with the SVM classification results from chapter 6 to enhance the accuracy of detecting instances where change points occur. 13 out of the 16 identified change points were appropriately corrected as a result of implementing this comparison. Though not contributing massively to the overall classification accuracy, this result demonstrates that a relatively simple step can enhance the accuracy of identifying modal change points within a GPS track significantly. The following chapter (8) describes the usage of transport network datasets performing a process we call “network matching” to verify modes of public transit in the results from the classification phase.

Chapter 8

Phase III: Verification (Network Matching)

8 PHASE III: VERIFICATION (NETWORK MATCHING)¹²

This chapter describes the last phase within the “mode of transport” detection framework from GPS data. This phase verifies whether the modes identified from the SVM classification (developed in chapter 6) follow their respective transportation networks. We assign this process the term of “Network Matching”. The networks involved in this verification stage are restricted to public transit networks (bus, train and underground) since they are of a unique and single-moded nature. Section 8.1 discusses the reason this test is conducted, while section 8.2 describes the available network data to this research and the reason of only matching to public transit networks and dismissing the rest. Section 8.3 describes a method that detects underground tunnel travel since no GPS fix data is available, and hence no network can be followed. On the other hand, sections 8.4, 8.5 and 8.6 describe the Network Matching tests for the three chosen networks; namely, the underground (non-tunnel), train and bus networks. The three sections also discuss the constraints and steps involved in the matching process along with the results of each type of matching. Section 8.7 describes how the Network Matching Test results are applied to the SVM modal classification, accounting once for each segment and once for each stage as an entity on their own. We conclude with section 8.8 presenting and discussing the results and accuracy achieved after the Network Matching stage is applied. The data used in this chapter for testing the Network Matching algorithm is the pilot data (21 users). The same test is applied to the 95-participants dataset in the following chapter (9) demonstrating the effect of adding participant-related information beforehand such as the ownership of a bike or a car, having a driving license, and access to local bike-rental services such as Barclays Bikes in London.

8.1 Why Matching to a Network?

As described in the chapter 2 (literature review), network matching, or “NM” as it will be referred to it in the rest of this chapter for short, is the process of assigning trajectory data to its appropriate transportation network. This must not be confused with map matching, which is the process of assigning every positional fix point to its corresponding network link in a given network, or as defined by Quddus (2006) as the process of snapping the displaced GPS points to the correct road link. On the other hand, NM is the process of selecting the rightful network on which a trajectory is travelling. An example is detecting that a group of GPS points representing a person travelling by train is travelling on a train network. This could be applicable to several networks such as bus, train, underground, road, cycling, etc.

The NM process can be used for several purposes, for example:

- As a primary step before map matching (To have a prior knowledge of which network to snap to)
- To analyse different network usages (or network split)
- To detect/verify the transport mode

¹² Most of the chapter is based on a previous publication of ours: Bolbol, A., Cheng, T. & Tsapakis, I., 2013. Matching GPS Data to Transport Networks. Geographic Information Science Research in the UK Conference, University of Liverpool, UK, April 3-5, 2013.

The latter is the very purpose of this research. In this chapter, we develop NM algorithms that will act as a verification phase of the initial transport mode classification from the SVM classification method developed in chapter 6. This stage is the final phase of the whole classification framework verifying if certain classified transport modes actually follow their corresponding network. For example, if a stage (a group of GPS segments) was classified as train, the NM phase will check whether all the GPS segments within that stage are within a pre-specified distance from the train network and follow a specific train route. This is illustrated in Figure 8.1 as two stages one of car mode and the other of train. The figure shows that the car fixes clearly follow the road network while the train fixes follow the train network.

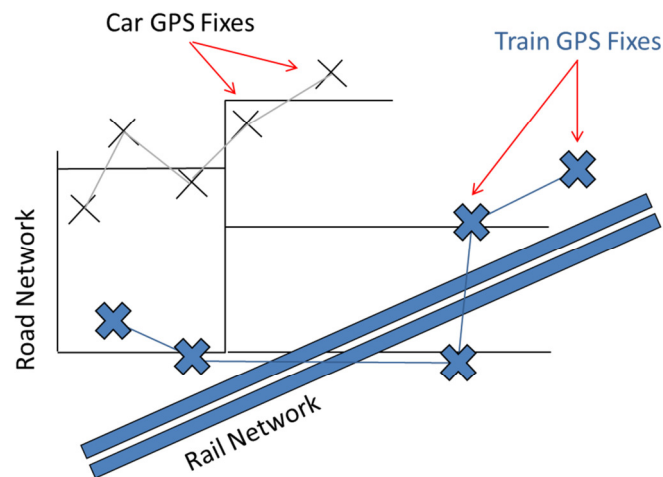


Figure 8.1 Conceptual Idea of Network Matching

In this work specifically, we are constrained to specific transport modes that exist within London, of which some can be network-matched and some cannot. The next section discusses the network types and the constraints that face the NM process for London.

8.2 Available Network Data

Many datasets are available to this research and could be used to enhance the accuracy and expand the capabilities of the developed framework. Among these datasets are the different transportation networks. London is a very transport-rich environment containing many forms of transportation infrastructure, and therefore it contains a huge number of different transportation networks. The network data available to this research is supplied by different sources (TfL and OSM) and their spatial accuracy might be subject to criticism; however we assume its validity for adoption to test the framework developed in this research. These different networks are illustrated in Figure 8.2 as the road, underground, train, footpath and bus networks.

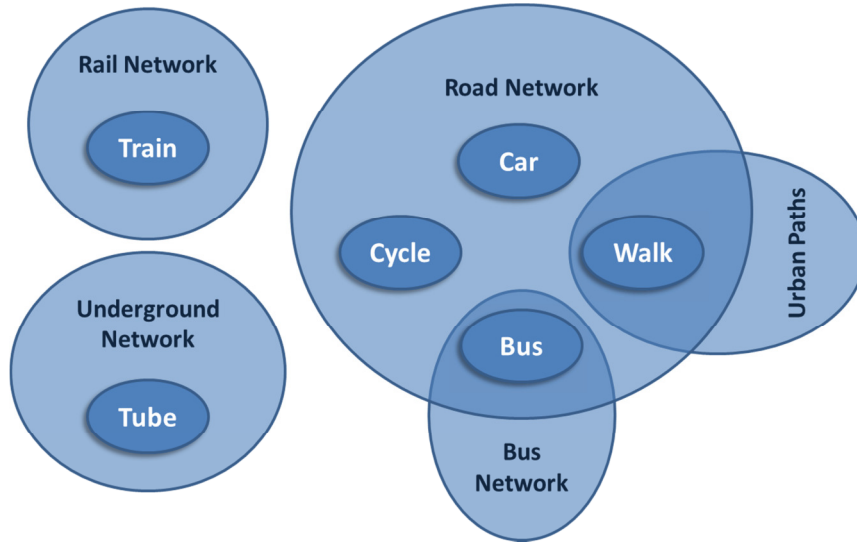


Figure 8.2 Different London Transportation Networks

Each network holds one or more travel mode type and vice-versa. The ITN road network for example holds cycle and car modes, while the train network only holds the train mode and the Urban Paths dataset (newly released by the Ordnance Survey in 2009) holds all the links that are only restricted to walk mode. This could be of great importance to the travel mode inference where any movement on the Urban Paths network infers a walk mode. However, the walk mode has been proven to be quite easy to detect and separate from other modes (as discussed in chapter 6). This eliminates the need to use the Urban Paths network in the planned NM verification phase.

Another element to bear in mind when applying NM is the time cost needed to run the applied algorithm, or the efficiency of the algorithm. In computer science, expressions such as the “big O notation” can be used to classify algorithms by how they respond (e.g., in their processing time or working space requirements) to changes in input size (Cormen, et al., 2009). Conceptually, in terms of this work, the size of the network (number of links) has a major impact on how fast any analysis or algorithm being run will take. This means that the size of the network is inversely proportional to the speed of a process such as NM. In other words, matching to a network such as the underground network will perform much faster than matching to the ITN (road) network considering the number of links. Therefore, for a faster performance, the NM algorithm should aim to avoid analysing huge networks unless necessary.

For this reason, we only select specific networks for the NM process. The simplest is the underground network, containing n polylines for n different underground lines. Also from the results of the SVM classification in chapter 6, the tube mode appears to be sometimes misclassified as car or bus. To address this, bearing in mind the computation cost, an easy way to enhance the classification accuracy is to apply NM initially to the underground network (described in section 8.4). However, as only 55% of the underground network is over-ground (TfL, 2012a), the NM only applies to this part of the network that is over-ground where GPS fixes are possible. Nevertheless, the underground tunnelled part is one of the cases where the absence of information is information in itself. The travelled segments in the underground

part of the network can be detected by the disappearance of a trajectory and reappearing at some other location, applying some constraints and having both locations being within a vicinity of an underground station. This will be described in detail in section 8.3.

The train mode has a similar situation as the underground network in terms of the network structure and the nature of trajectories travelling on it except for being totally over-ground and being a larger dataset covering the whole of the UK. As noted from chapter 6, the train mode is often confused as car or bus modes. Similar to the underground mode, we will apply NM to the train network to resolve this confusion. This is discussed in section 8.5.

Similarly, the bus mode is sometimes misclassified as car or cycle. This could also be addressed by applying NM to the bus network. Given the full bus network data provided by TfL, a NM algorithm could be applied to verify bus classifications by testing whether a trajectory is following a specific bus route or not. We describe this process along with its constraints in section 8.6.

Finally, the car and cycle modes are the only two non-walk modes that will be left without NM. As previously demonstrated in chapter 6, the confusion between both modes is immensely lower when classified using acceleration. Therefore, the confusion between both modes can be resolved using results obtained from the SVM classification from acceleration data without any NM.

Section 8.7 demonstrates how the verification process and final classification is decided upon in cases of contradicting evidence of trajectories being matched to several networks. On the other hand, section 8.8 then discusses the NM accuracy attained for each network type bearing in mind network types and the sample size for each mode within the testing dataset.

8.3 Underground (Tunnel) Tube Network Matching (Time-Distance Underground Travel Detection Algorithm)

This test aims at detecting the underground (tunnel) segments of the underground travel mode within a GPS track. The test largely depends on the nature of that specific type of trajectory. This might change from one tube network in one city to another. The complexity of the London underground (tube) network makes it difficult to differentiate a trajectory's location as either being on the tube or the road network. On the other hand, its complexity sometimes makes it easy to detect the respective network on which a trajectory is travelling due to its distinctive movement patterns. This section describes the method we developed to detect underground travel using the London Underground network. The method relies mainly on the underground stations point dataset since modal switches into underground tunnel movements usually appear and disappear at locations of these stations. The following subsection (8.3.1) describes the network properties and restrictions, while the remainder of the section describes the adopted method in details along with the accuracies attained upon testing it.

8.3.1 Network

The London Underground network consists of 270 stations and 402 kilometres (250 mi) of track (TfL, 2012a), making it the second longest metro system in the world after the Shanghai Metro. It also has one of the highest numbers of stations. In 2007, more than one billion

passenger journeys were recorded, making it the third busiest metro system in Europe after Paris and Moscow.

This massive network's data resource is provided to this research by TfL. The data provided contains all the underground stations and lines. Each line (e.g. District, Northern, etc.) is held within a unique polyline with a unique identifier. An illustration of the underground network is shown in Figure 8.3 along with the over-ground and national rail networks in the background. Figure 8.4 shows the underground stations of the London Underground network superimposed on the rest of the network.

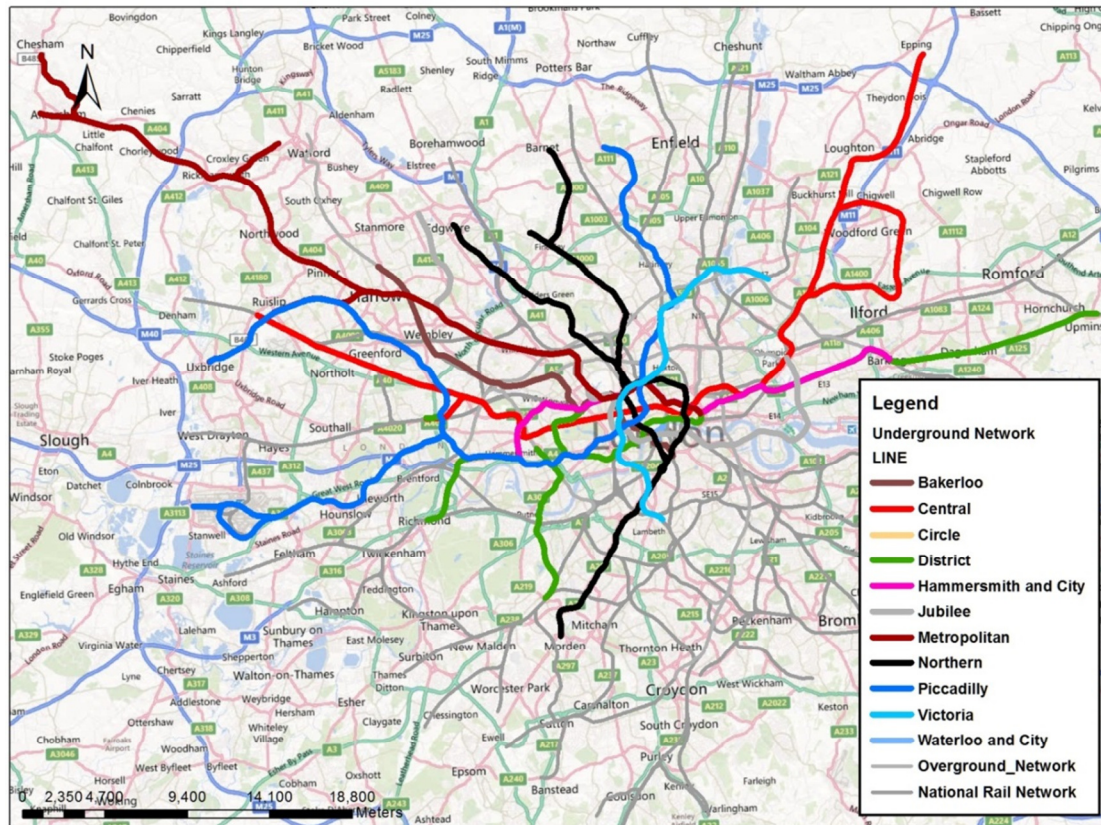


Figure 8.3 TfL's London Underground Network Data (Data provided by: TfL)

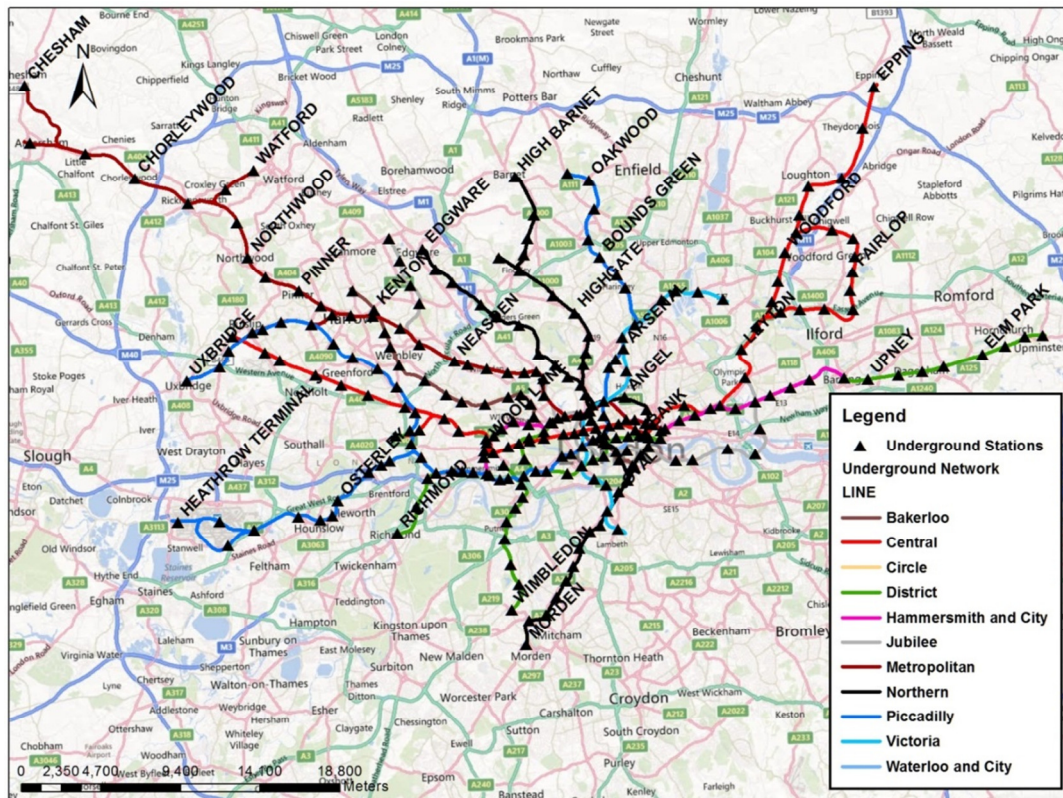


Figure 8.4 Tfl's London Underground Network & Stations (Data provided by: Tfl)

Some of the characteristics of an underground trajectory could include the following:

1. Around 45% of the tube network is underground, however, 75% of the trips take place within the central zones where the network is underground, which means that in many of the cases there will not be any GPS fixes attainable for the trips.
2. All 100% of the tube trajectories follow the underground network.
3. The distance covered by the tube in most journeys is far faster than could be covered by any other form of commute in an urban area (discussed in section 8.3.3).
4. The time spent on the underground network most probably would not exceed a certain temporal threshold (discussed in section 8.3.3).
5. Any underground journey starts with an entrance and exit to and from a tube station (discussed in section 8.3.3).

In the following subsection, we use these characteristics to identify the occurrence of underground travel.

8.3.2 Principle

For over-ground tube trajectories; the network matching test (section 8.4) verifies their classification by matching them to the Underground network. It also verifies whether the trajectory starts and ends the stage at underground stations. Whereas for underground trajectories; we use the loss of data (loss of GPS signal) as information in itself. The idea simply is to find instances when there was a long distance travel without GPS coverage in a relatively short duration of time. This does not ignore the spatial domain but it only applies to these gaps that start and end at an underground station.

The box plots presented in Figure 8.5a and Figure 8.5b illustrate the distance and speed data of the pilot database respectively for the underground trips. The data is grouped separately as for time difference less than 2 minutes and for more than 2 minutes (twice the epoch rate) for both plots of distance and speed. It can be evidently noted that for data less than two minutes the distance is relatively small compared with data longer than 2 minutes. This demonstrates that when there is constant fix capture the travel is for short distances, while when there is a loss in fix more distance is covered. Although this is quite evident, the speed data presents a different case. From Figure 8.5b, the speed for less than 2 minutes is much higher than when there is a loss of fix which is probably due to underground travel. Analysing this phenomena, we can deduce that speed of underground travel is slower than over-ground even if the distance covered is larger, which is probably due to the time taken for a participant to walk into or out of the underground station, or to wait for the next service to reach the platform and maybe other activities such as purchasing a ticket or topping-up Oyster credit (credit card for London Underground usage). The different nature of the underground travel segments makes them difficult to be detected using SVM classification from speed or acceleration data. On the other hand, as demonstrated from Figure 8.5a, distance travelled for a range of time intervals can be used to detect this kind of movement. Hence, the thresholds of time and distance to detect underground tube travel need to be set in order not to under- or over-classify underground tube mode travel.

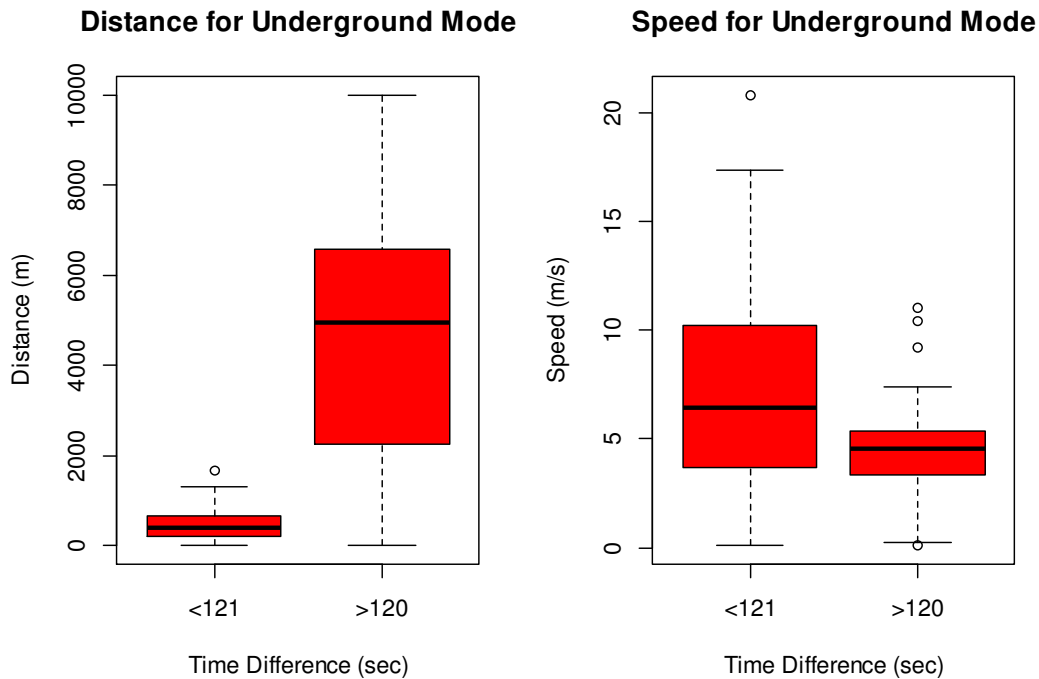


Figure 8.5 Distance(a) & Speed(b) travelled for Segments of less & more than 2 Minutes in Underground Mode

8.3.3 Constraints

Thus, the three constraints for this test are; (a) **time** spent travelling, (b) **distance** travelled and (c) **distance from tube stations**. For these three constraints to be controlled, we need to set upper and lower thresholds for each of them.

For the **time** constraint, upper and lower thresholds have to be set in order to constrain the detection of underground tube travel within logical travel times. The time taken to travel underground (in the central zone) between two stations is 2-3 minutes on average plus assuming an extra 2-3 minutes for stoppage time and to start moving again at an underground station (TfL, 2012c). This makes a 5-6 minutes loss of fix time in the underground, and therefore, we can set a threshold of 5 minutes or **300 seconds** for the minimum time for which we can consider a segment for entering the time-distance algorithm.

A passenger in the London Tube Underground can be underground for as long as they wish, however, a crude assumption must be made about the maximum duration one can spend underground during a tube journey. Given that longest journey without changing trains – 59.4 kilometres (36.9 mi), from West Ruislip to Epping on the Central line takes 1 hour 28 minutes (around 5400 seconds) (TfL, 2012a), we can set this as our maximum temporal threshold. Moreover, Figure 8.6, which plots the time versus distance for the underground mode from the pilot data, shows that most of the data with long time difference (i.e. when loss of fix occurs) rarely exceeds the 5400 seconds mark. This suggests that the upper threshold of **5400 seconds** could serve as a valid assumption.

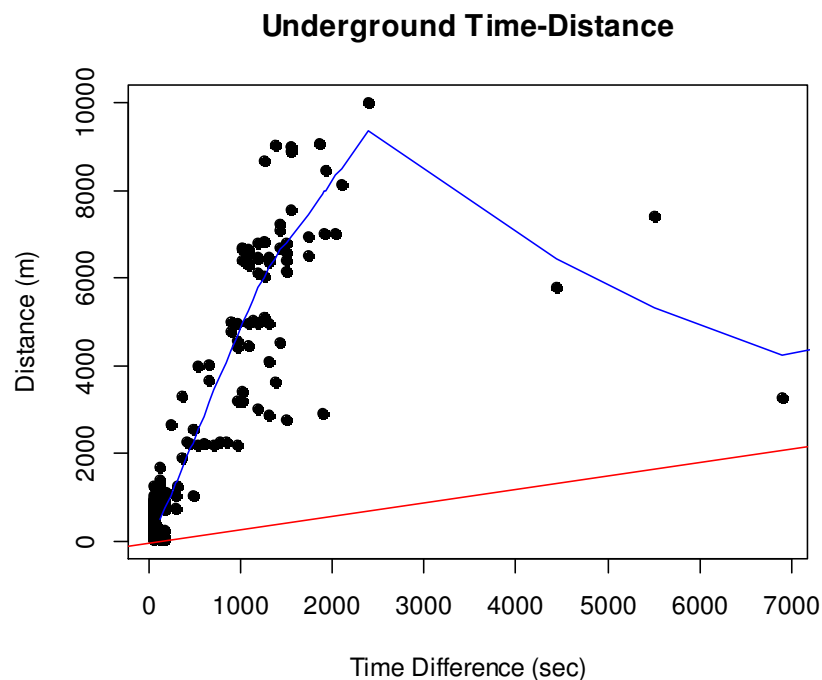


Figure 8.6 Time-Distance Scatter Plot for Segments of Underground Mode

As for **distance**, as can be noted from the figure 8.5 a, the upper and lower thresholds of the distance travelled underground are around 2000 and 10000 meters. The lower distance threshold could be taken as **2000 meters** or less as the other constraints will rule out any over-classification. Nevertheless, as stated by TfL (2012a), the upper threshold could be set as

the longest continuous tunnel: East Finchley to Morden (via Bank) - 27.8km/17.25 miles (or **28000 meters**).

As for the *GPS points distance to the underground stations* upon entering into, exiting from or appearing in-route at a station, we need to consider two factors. The *first factor* is the last or first fix before or exiting an underground station. An evident assumption is that a passenger will enter into/exit a station walking. Therefore, applying a normally distributed average speed of 1.34 m/s and a standard deviation of 0.26 for an average pedestrian walking on the street (Bloomberg & Burden, 2006), a distance of around 80 meters can be covered within one minute which is the epoch rate for data collected in this research. As for our threshold, we can take double this distance (160 meters) to avoid any chance of an occurrence of a loss of signal once before entering or after exiting the station. The *second factor* is the GPS device accuracy. Therefore, we also add an allowance of 50 meters, which is the twenty times the highest error level expected by GTrek devices (GTrek, 2012) used for the validation dataset of 95 participants or 10 times the u-blox devices (u-blox, 2009) used for the pilot dataset. Also, accounting for sources of GPS error such as multipath in urban canyons, the final threshold for the distance from underground stations is $((2*80)+ 50) = \mathbf{230\ meters}$.

As a result, for the time threshold we will use 300 and 5400 seconds as lower and upper thresholds respectively, and similarly 2000 and 28000 meters for distance. As for the distance from underground stations, we use a threshold distance of 230 meters. The following subsection describes the different steps that use these thresholds to detect the underground tube mode of transport.

8.3.4 Algorithm and Limitations

The algorithm is developed in R Project for Statistical Computing platform using different relevant libraries such as for example; “fields” to find the shortest distances to underground stations (R Project, 2012). The algorithm first runs row-by-row through every segment to query for the appointed time-distance thresholds. Once a segment is selected, both ends of the segment are matched to the nearest tube stations. This is achieved using an R Project library named “Fields”. The “fields” library enables dealing with spatial data such as calculating the distances between points in 2 datasets, outputting a matrix of distances between every pair of coordinates in one set and every pair in the other. Using this function, a matrix of distances between each end of the segment and all tube stations is calculated. Then the shortest distance between the each end of the selected segment and all the tube stations is selected, outputting the nearest station to each end (e.g. “Bayswater” as demonstrated in Figure 8.7). A check then is applied to whether the distance between each end and the nearest station is less than the appointed threshold (230 meters). If it is not below the threshold, the algorithm discards the segment and moves on to the next segment.

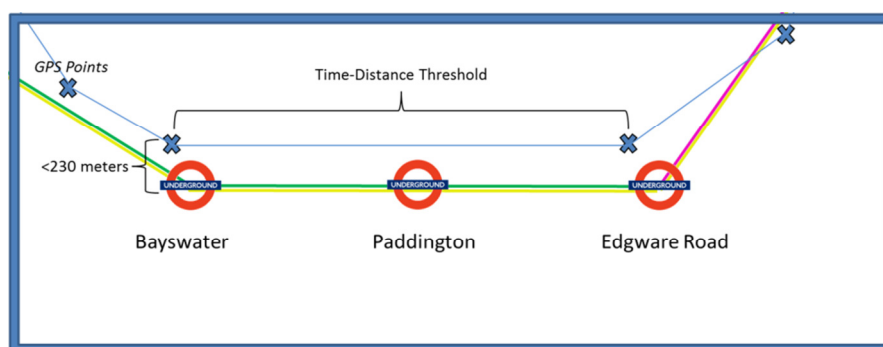


Figure 8.7 Description of Underground Tunnel Detection Algorithm

There are two limitations, however, that might arise from this process. The **first** of which is that the GPS battery could have run out and the participant turned it on later on in the track. This means that the algorithm will pick these instances up if a distance less than 27 km has been travelled with no fix and the device was turned on again before an hour and a half have passed. This can lead to over-classifications. Hence, the upper threshold of distance will be better decided upon after the final results of the 95 participants are tested in chapter 9.

The **second** limitation is that either ends of an underground segment need not to be at an underground station at instances when the tube comes out of a tunnel on a given underground route in the middle between two stations. This means that these instances would not be picked up by the algorithm leading to an under-classification. This is later resolved by applying NM to the underground network as both ends of the segment will be matched to the network, and therefore, classified correctly as an underground mode.

8.3.5 Matching Accuracy

The accuracy achieved by this method was 66 segments out of a total of 319 underground segments (which are not necessarily of the tunnel type), and achieving a precision of nearly 90%. That is, apart from all the detected tube modes (90%), other modes were also misclassified as underground (10%). This is illustrated in Table 8.1 showing the confusion matrix of the classification resulting from this method for the pilot data. As might be noted, there is some confusion with the bus mode (4 instances), which is considered very minimal.

| Truth | Inferred | |
|-------|------------|--------|
| | Not Tunnel | Tunnel |
| Bus | 1212 | 4 |
| Car | 2208 | 2 |
| Cycle | 1456 | 0 |
| Train | 282 | 0 |
| Tube | 253 | 66 |
| Walk | 10355 | 2 |

Table 8.1 Time-Distance Underground Tunnel Travel Detection Method Confusion Matrix Results

The time-distance algorithm produces an output CSV file with nearest station names and distances to them from each end of the segments under investigation. Table 8.2 shows an example of this output. The “Station Match” column shows the final classification after verifying whether both ends were within 230 meters from an underground station.

| Seg. Id | E1 | N1 | E2 | N2 | Length (m) | TimeDiff (s) | Distance1 | NearestStation1 | Distance2 | NearestStation2 | Station Match |
|---------|----|----|----|----|------------|--------------|-----------|------------------------------|-----------|--------------------------------|---------------|
| 141 | - | - | - | - | 2261.77 | 420 | 87.23 | Hammersmith Edgware Road (B) | 412.29 | Marble Arch | NA |
| 981 | - | - | - | - | 2207.33 | 454 | 93.34 | | 90.84 | Euston Square Edgware Road (B) | tube |
| 1049 | - | - | - | - | 2196.86 | 718 | 32.34 | Euston Square | 67.54 | | tube |
| 1474 | - | - | - | - | 2893.54 | 1900 | 349.96 | Charing Cross | 295.06 | Barbican | NA |
| 1527 | - | - | - | - | 8477.14 | 1923 | 56.49 | Paddington | 45.33 | Hammersmith Edgware Road (B) | tube |
| 2340 | - | - | - | - | 1238.87 | 307 | 102.77 | Marylebone | 88.21 | | tube |
| 2341 | - | - | - | - | 2181.26 | 541 | 23.52 | Paddington King's Cross St. | 111.07 | Euston Square | tube |
| 112 | - | - | - | - | 3432.09 | 1020 | 107.11 | Pancras | 228.29 | Charing Cross | tube |
| 126 | - | - | - | - | 3624.75 | 1380 | 20.45 | Paddington | 20.45 | Paddington | tube |
| 172 | - | - | - | - | 1529.32 | 420 | 80.77 | Victoria | 77.74 | Marylebone | tube |
| 501 | - | - | - | - | 1142.64 | 3419 | 156.92 | Regent's Park | 225.45 | Euston Square King's Cross St. | tube |
| 1051 | - | - | - | - | 3813.14 | 361 | 290.67 | Colliers Wood | 314.62 | Pancras | NA |
| 1052 | - | - | - | - | 4089.44 | 419 | 65.65 | Bank | 144.01 | Paddington | tube |

Table 8.2 Example of Output File of Time-Distance Underground Travel Detection Algorithm

Figure 8.8 shows all trip segments that satisfy the time-distance threshold constraints plotted on a map of London with the Underground lines and stations in the background. As can be noted on the map, most segments that start and end near underground stations (in red) are correctly detected as underground tube trips. On the other hand, trips that do not start or that do not end near underground stations are not classified as underground tube trips (in grey). This is more clearly visible in Figure 8.9 which is a smaller scale map of the same classification. The unmatched segments (in grey) seem to be just random occurrences of the device being switched off and on again, or just an unexplained loss of fix, but not underground travel.

8 Phase III: Verification (Network Matching)

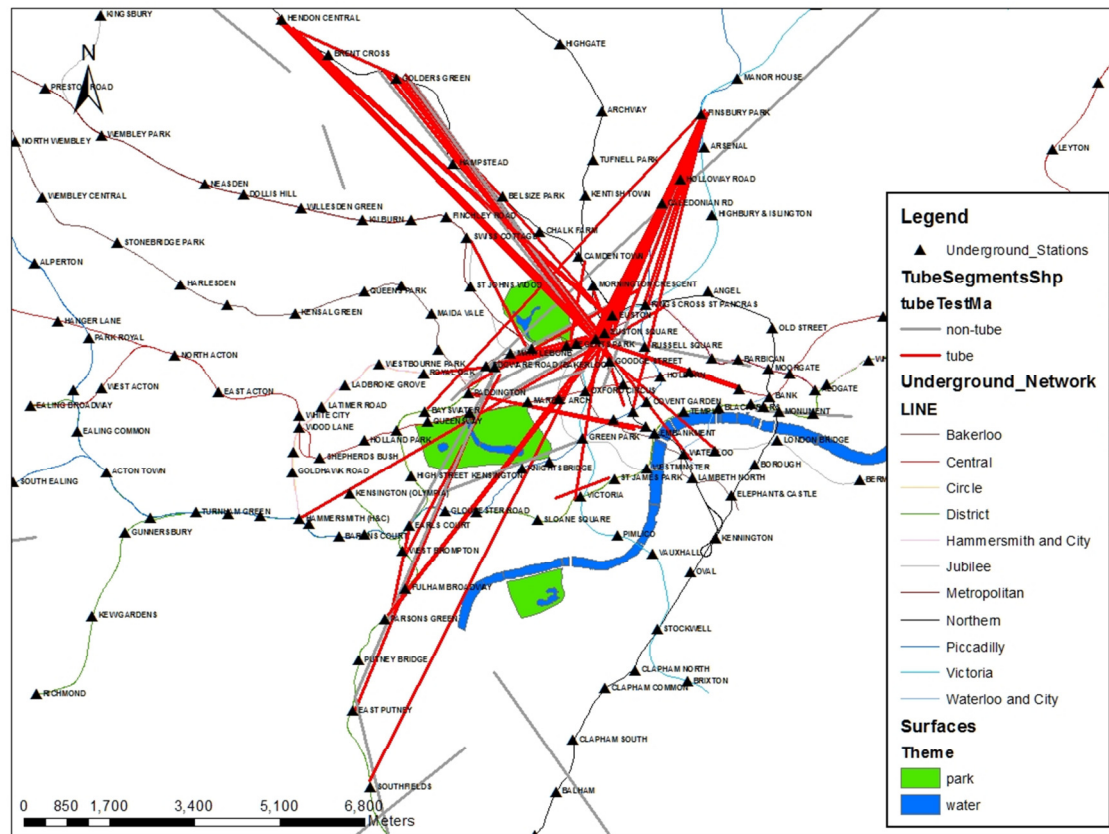


Figure 8.8 Underground Travel Detection Results for Pilot Dataset

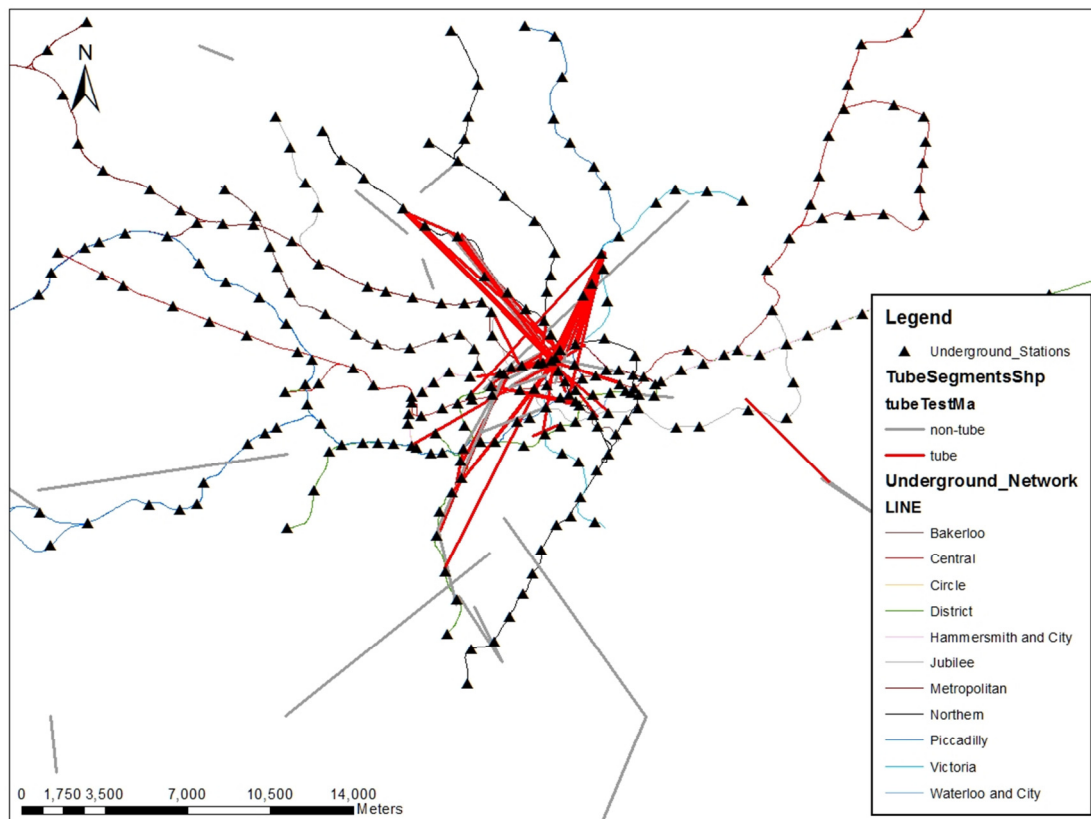


Figure 8.9 Underground Travel Detection Results for Pilot Dataset (Smaller Scale)

The results of the 95 participants' dataset are demonstrated in chapter 9 in full detail analysing the zones where underground travel occurs most within the collected dataset. The next section, however, describes the NM algorithm for tube usage in the case of existence of GPS fixes on the part of the underground network that is over-ground (not in tunnels).

8.4 Underground (Non-Tunnel) Tube Network Matching

The total number of passengers carried each year by the London Underground network is around 1,107 million (TfL, 2012a). This shows how much the tube mode contributes to the modal mix for an average Londoner. The total length of the network is around 402 kilometres (249 miles) covering most of the congested London boroughs. These facts motivate this research to accurately account for tube travel and to being able to accurately classify tube mode occurrences.

The test developed in this section aims at detecting whether a non-walk trajectory follows the underground network, leading to the verification or amendment of results from the SVM classification conducted in chapter 6. The algorithm is developed in R Project for the main analysis and in Python for the usage of ArcGIS functions (ArcGIS, 2012).

8.4.1 Network

As demonstrated in the previous section, the underground data is all provided by TfL. Other datasets exist such as user-generated OpenStreetMap online data for the same dataset (OSM, 2012). Table 8.3 shows some of the differences between both network datasets. The main problem with the OSM dataset is that each underground line does not contain a unique identifier, while for the TfL dataset, each line (e.g. District, Northern, etc.) is held within a unique polyline with a unique identifier. This makes the matching process more accurate, since the algorithm can query whether a trajectory is following the same underground line and not jumping from one line to another. As a result, we run this test using the TfL network dataset.

| | TfL National Rail Network | OSM National Rail Network |
|--------------------------|---|--|
| Accuracy | Highly accurate (TfL accuracy) | Variable accuracy (User-collected) |
| Coverage | London area only | Nationally all UK |
| Naming | Very accurate and uniform | Accurate but not uniform |
| Unique Identifier | Data of one route form a single polyline that has a unique identifier | Each route consists of many polylines that are not related using a unique identifier |

Table 8.3 Differences between Network Data provided by TfL & OSM

As demonstrated in subsection 8.3.1, the London Underground contains 11 lines (routes). The number of km/miles travelled by each Tube train each year is around 184,269 kilometres (114,500 miles) (TfL, 2012a). The average train speed is around 33 kilometres per hour (20.5mph). This speed falls within the range of speeds of many other modes which easily can lead to the misclassification of underground mode segments into other modes. This motivates this test in order to verify the classification of underground mode or to rectify any misclassifications into other modes.

8 Phase III: Verification (Network Matching)

8.4.2 Principle

The over-ground tube NM algorithm (or a NM algorithm in general) *aims at verifying whether a stage of a given non-walk mode follows the underground network in a unique route*. Even if a participant uses the underground service and switches from one route to another, a stop/or walk mode will interrupt the stage, and hence, the trip will constitute of one tube stage, stop/or walk, then a tube stage again.

The NM algorithm is similar for each of the bus, train or tube networks except for minor differences due to the different nature of the network, its dataset structure or the nature of movement of the trajectories that use it. The algorithm starts by finding the nearest n underground routes ranked according to how close each route is from a GPS fix from closest to furthest, and assigning this information into a matrix. The GPS fixes of any non-walk stage are then tested to whether they all fall within a threshold distance from the same underground route. The duration of the stage is also constrained to a temporal threshold otherwise the stage is discarded from being underground-network-matched, in order to avoid pauses in the track such as a 1-day gap in the GPS track. If a given stage is matched to the underground network, it is then noted as such, as well as the name of the nearest route. Once that stage is tested for bus and train NM too, all the results are queried at a later stage (described in section 8.7) and a final classification is assigned to the stage.

8.4.3 Constraints

There are two constraints for this test, namely; (a) the **maximum distance** allowed to the underground network to allow the inclusion of an underground route, and (b) the **minimum time** duration threshold that a given stage is allowed to be, for it to be qualified as a candidate for being matched to the underground network. These two constraints are demonstrated in Figure 8.10 as an underground segment constrained by these two thresholds. These constraints will exist for all the public transport networks; however, their threshold values will depend on the nature of movements of the trajectories on each network and the nature network structure itself.

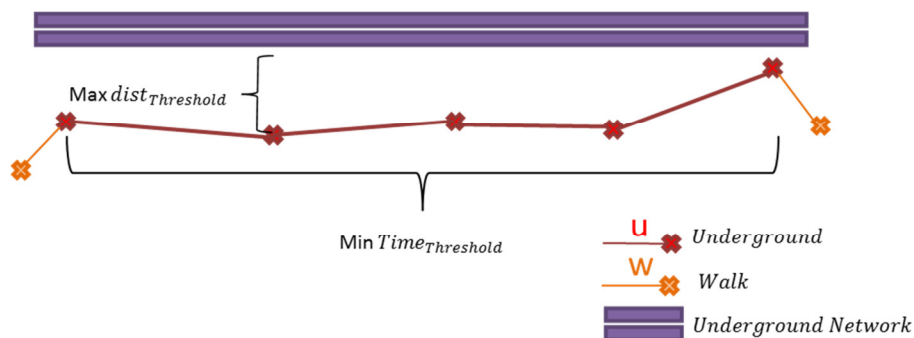


Figure 8.10 Tube Network Matching constrained by Temporal & Distance Thresholds

As for the **distance** from the underground routes, it will depend on the GPS accuracy and the accuracy of the map source. The map source is assumed to be trustworthy as provided by TfL leaving us solely with the GPS accuracy level. The stated positional accuracy of the GTrek devices is ± 3 meters (GTrek, 2012) for the validation dataset and ± 4.3 meters for u-blox devices (u-blox, 2009) for the pilot dataset. Nevertheless, due to the nature of the data being collected

in an urban environment, many sources of error affect this accuracy such as multipath, bad GDoP, and Ionosphere and Troposphere disturbances (Hinch, 2007). Yet, for the purpose of this algorithm, a generous distance is more plausible since the test will check if all the consequent points will also fall within this specified distance threshold, which provides another checkpoint for the classification. Therefore, we assign a distance of **150 meters** which is around 30 times the stated positional accuracy of the u-blox devices to account for the extreme urban canyon effect.

On the other hand, for the **temporal** threshold we need to assign the minimum journey time of a tube stage that occurs above the ground since below-the-ground sections of these trips are identified using the time-distance method (section 8.3). According to TfL (2008), the shortest network distance between two stations is between Covent Garden and Charing Cross (250m), a section which is covered below the ground. On the other hand, the shortest network distance above the ground is between South Ealing and Northfields (380m) which also happens to have the shortest travel time for the part of the network above the ground (1.5 minutes). This minimum travel time between two tube stations accounts for train dwell time which includes time for passengers to get on and alight from the train carriages (TfL, 2008). This 1.5 minutes value is close to the epoch rate set for the GPS data collection in this research. Hence, we do not need to assign a minimum threshold for the time spent on an underground stage trip.

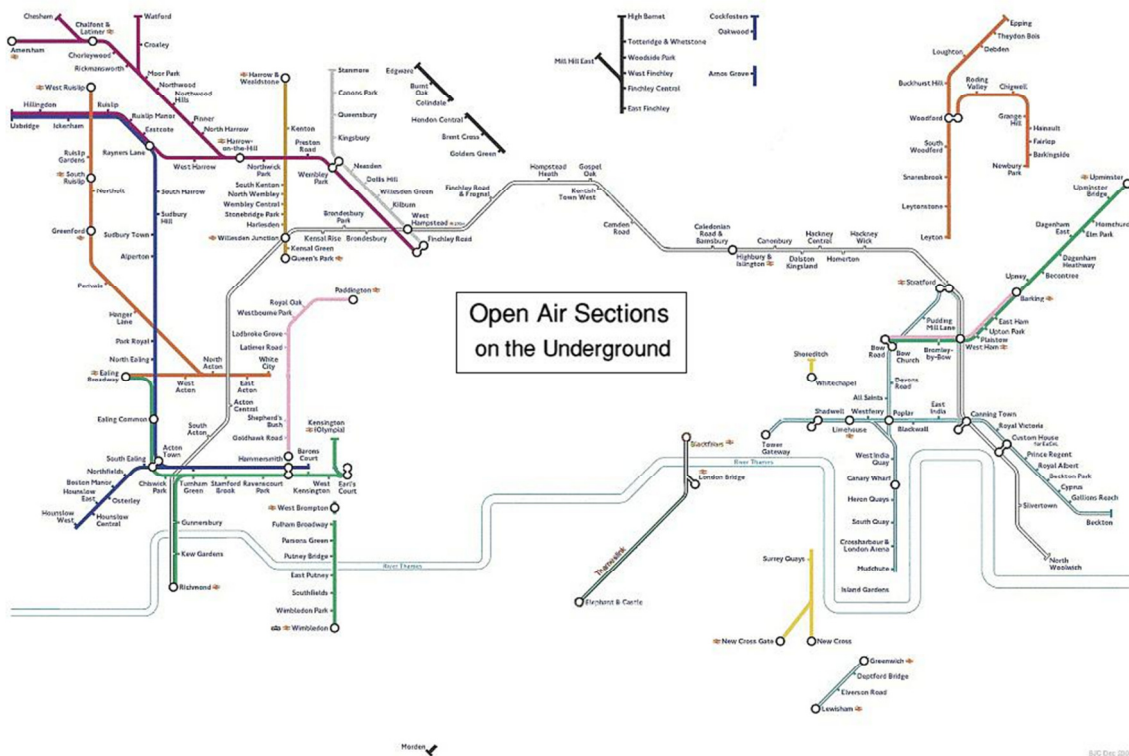


Figure 8.11 London Underground Tube Sections appearing above Ground (TfL, 2012d)

8.4.4 Algorithm and Limitations

The NM algorithm for the underground network is divided into two phases; a phase developed in **Python** and another developed in **R project**. The Python phase is run before the R Project phase in order for the former to generate a distance matrix to be used in the latter to assess

the vicinity of the GPS fixes to the underground network. The algorithm is summarized in Figure 8.12 as a flow chart and is described afterwards.

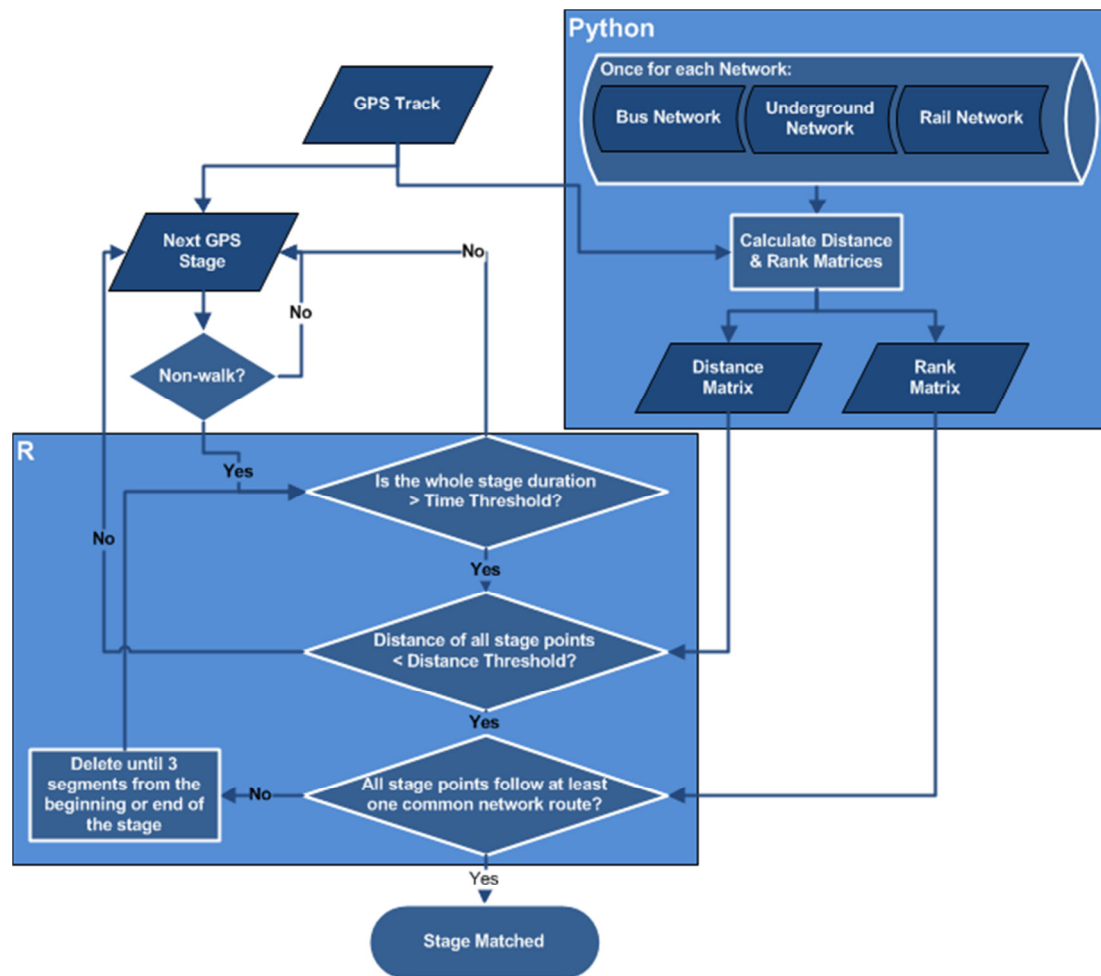


Figure 8.12 Network Matching Algorithm Flow Chart

Phase 1:

1. First, in Python command line using ArcGIS's library "ArcPy" (ArcGIS, 2012), we use the `GenerateNearTable_analysis` and `PivotTable_management` functions which generate a "Near Table". The "Near Table" is a table which holds each GPS fix and the nearest tube lines ranked according to how close they are to a GPS point from nearest to furthest. Since there are 11 underground lines (routes), then the near table will hold 11 columns, each containing an underground line.
2. There are two outputs from this process: (a) A **rank** table: holding the names of each of the underground lines ranked from closest to the furthest for each GPS fix. (b) A **distance** table: listing the distances to each of the underground lines listed in the rank table.

Phase 2:

3. In R project, every non-walk stage of either of cycle, bus, car or train is selected to enter into the NM algorithm.
4. The two outputs from the python code are then read into R Project, namely; the rank and distance tables.

5. Each row (GPS fix) in the distance matrix is then tested to whether it is greater than the assigned threshold (150 meters), and if so its corresponding underground line from the rank table is discarded. Now, the ranking table is left with lines that are maximum 150 meters away from each GPS fix.
6. The matching process then begins as follows:
 - a. Each non-walk stage is tested to whether it is longer than the given threshold of time (5 minutes), and is discarded if it is shorter than the threshold.
 - b. In the base library or R Project, the “intersect” function searches if there are any common underground routes that are within the distance threshold (150 meters) along all the stage. This essentially means that this phase of the algorithm checks if the whole stage follows one underground route or not. Figure 8.13 shows an example of all the possible matches to the underground network from different classified modes.

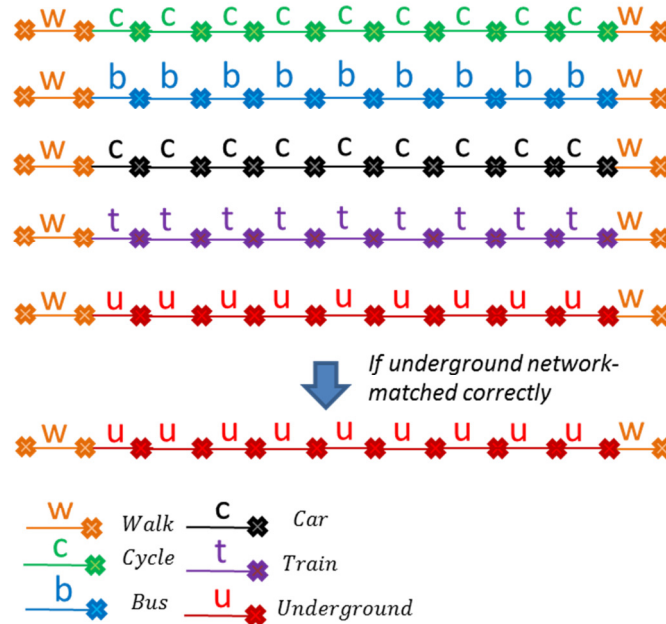


Figure 8.13 SVM Classified Segments used for Public Transport Network Matching

- c. If a common matched underground route is not found, the algorithm repeats this process iterating through the same segment excluding one of the first or last 3 segments of that stage (to eliminate the chance that the one of these segments were included erroneously). This is demonstrated in Figure 8.14 showing the last two segments of the underground stage being dismissed and returned to walk mode, as they do not follow the same (if not any) underground route.

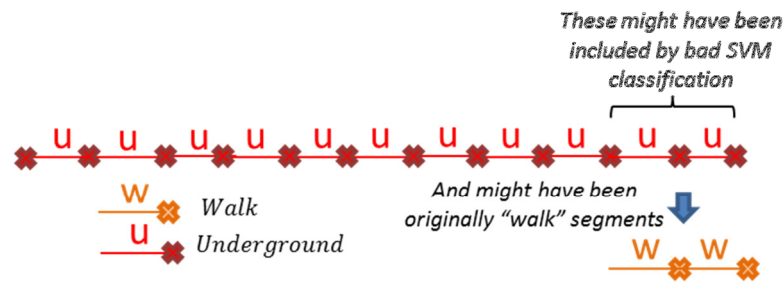


Figure 8.14 SVM Over-Classified Segments reversed by the Algorithm to Walk Mode

- d. If a stage is matched, the nearest underground lines within the threshold (150 meters) are noted. This is later used to decide upon the final classification after the bus and train NM are also conducted as described later in section 8.7 in detail.

8.4.5 Matching Accuracy

The underground NM process is a complementary process to the rest of the classification framework as it builds up on the initial SVM classification. Hence, we will assess the accuracy of the process not only on its own, but on how well it complements the initial SVM classification. Figure 8.15 shows an example of the matching to the “District” Line of the underground network (matched GPS fixes are denoted in red). Figure 8.16 also shows an example of repetitive trips matched to the “hammersmith”, “District” and “Circle” lines.

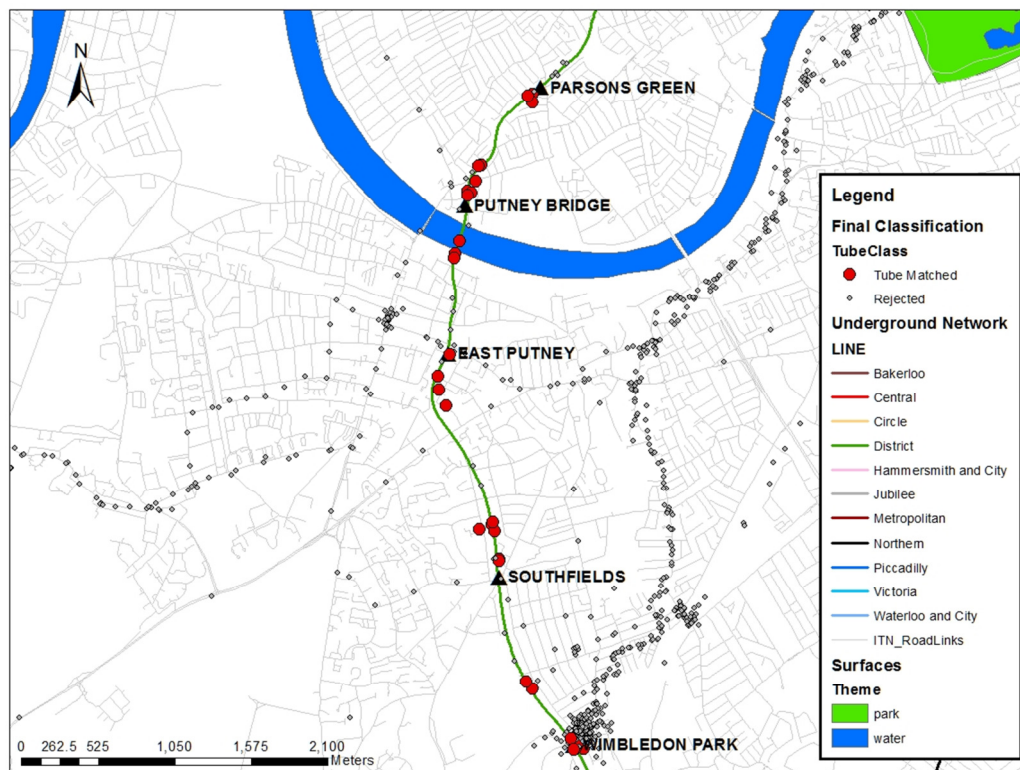


Figure 8.15 Example of Network-Matched GPS Fixes to London Underground Network (District Line)

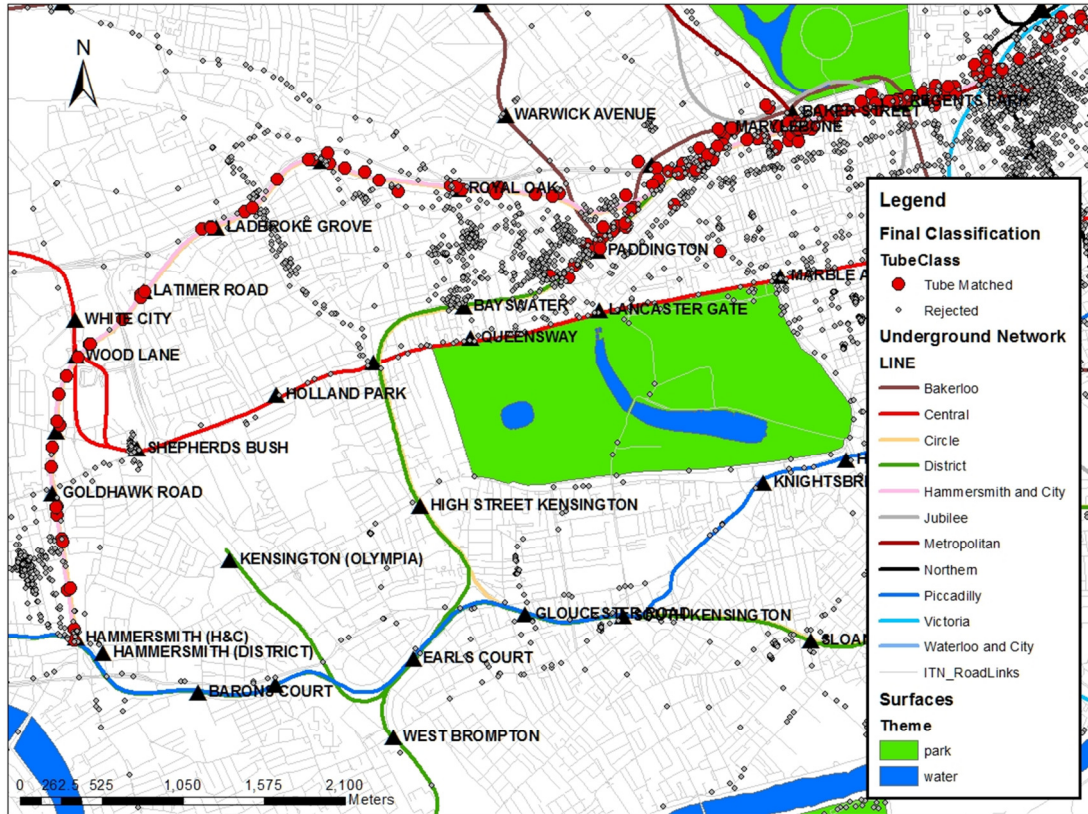


Figure 8.16 Example of Network-Matched GPS Fixes to London Underground Network for Repetitive Trips

Table 8.4 shows the results from the underground NM process for all tested segments from the pilot dataset based on the raw data and not building up on the SVM classification results and assuming a pre-defined segmentation. The purpose of the table is to quantify the accuracy of the algorithm in absolute terms without any dependency on the accuracy of the performance of the SVM classification stage. The table presents results as three categories. The “No Entry” category shows numbers of the GPS fixes that were not granted entry into the algorithm from each mode of transport. The “Not Tube” category shows the number of fixes that were declined from being matched from the underground network, while the “Tube” category shows the number of fixes that were approved as matched to the network. The table shows that more than two-thirds of the tube fixes were correctly matched ($\approx 69\%$). However, as described in the previous section, the long underground travel segments might not be matched to the network. Therefore, we show the same results in Table 8.5 dismissing these long segments that could already be identified using the time-distance algorithm, revealing a tube mode matching of 78%.

| Truth | Inferred | | | |
|-------|----------|----------|------|--------------------------|
| | No Entry | Not Tube | Tube | % of Tube Matched Stages |
| bus | 57 | 934 | 225 | 18.50% |
| car | 124 | 2048 | 38 | 1.72% |
| cycle | 47 | 1384 | 25 | 1.72% |
| train | 17 | 265 | 0 | 0.00% |
| tube | 40 | 59 | 220 | 68.97% |
| walks | 10357 | 0 | 0 | 0.00% |

Table 8.4 Underground Network Matching of GPS Fixes Confusion Matrix Results

| Truth | Inferred | | | |
|-------|----------|----------|------|--------------------------|
| | No Entry | Not Tube | Tube | % of Tube Matched Stages |
| bus | 57 | 931 | 224 | 18.48% |
| car | 124 | 2046 | 38 | 1.72% |
| cycle | 47 | 1384 | 25 | 1.72% |
| train | 17 | 265 | 0 | 0.00% |
| tube | 5 | 51 | 197 | 77.87% |
| walks | 10355 | 0 | 0 | 0.00% |

Table 8.5 Underground Network Matching of GPS Fixes Confusion Matrix Results (Dismissing the Time-Distance Outcome Trips)

In order to have a better quantification of the accuracy per number of trips, Table 8.6 shows the same figures shown in the previous two tables by the number of stages. The table shows that **68%** of the underground trips were matched correctly to the underground network. It might be noted, however, that another 32 bus trips (21% of bus trips) were incorrectly matched to the underground network. This is due to the fact that some bus routes follow closely underground routes. This conflict can be resolved when the bus NM is applied and further reasoning can be implemented. This is explained in section 8.7 while reasoning between the different results from the SVM classification and the NM of the different types of networks.

| Truth | Inferred | | | |
|-------|----------|----------|------|--------------------------|
| | No Entry | Not Tube | Tube | % of Tube Matched Stages |
| bus | 5 | 115 | 32 | 21.05% |
| car | 12 | 188 | 10 | 4.76% |
| cycle | 0 | 87 | 3 | 3.33% |
| train | 1 | 24 | 0 | 0.00% |
| tube | 5 | 8 | 27 | 67.50% |
| walks | 579 | 0 | 0 | 0.00% |

Table 8.6 Underground Network Matching of GPS Stages Confusion Matrix Results

In order to assess the classification algorithm developed in this research as a whole, we also need to calculate the accuracy of the NM performance based on the results from the SVM classification process. Table 8.7 shows the results of the underground NM, as a verification phase of the classified non-walk GPS fixes resulting from the SVM classification. Table 8.8 also shows the same results but as number of entire stages rather than GPS fixes. What could be noted is that 9 out of 40 (**22.5%**) of underground trips were correctly matched to the

underground network. This is due to wrong inclusion of some GPS fixes to some underground segments that do not belong to the stage. That is due to some misclassification caused by the SVM classification phase. Some bus trips are also matched to the underground network, but as previously mentioned, this will be amended after all the bus network matching is applied.

| Truth | Inferred | | | |
|-------|----------|----------|------|--------------------------|
| | No Entry | Not Tube | Tube | % of Tube Matched Stages |
| bus | 188 | 907 | 117 | 9.65% |
| car | 208 | 2000 | 0 | 0.00% |
| cycle | 74 | 1376 | 6 | 0.41% |
| train | 13 | 269 | 0 | 0.00% |
| tube | 59 | 106 | 88 | 34.78% |
| walks | 9431 | 891 | 33 | 0.32% |

Table 8.7 Underground Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results

| Truth | Inferred | | | |
|-------|----------|----------|------|--------------------------|
| | No Entry | Not Tube | Tube | % of Tube Matched Stages |
| bus | 42 | 98 | 12 | 7.89% |
| car | 49 | 161 | 0 | 0.00% |
| cycle | 17 | 73 | 0 | 0.00% |
| train | 1 | 24 | 0 | 0.00% |
| tube | 15 | 16 | 9 | 22.50% |
| walks | 397 | 172 | 10 | 1.73% |

Table 8.8 Underground Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results (Dismissing the Time-Distance Outcome Trips)

8.5 Train Network Matching

Although the modal share of the rail mode for Greater London is only 4% (TfL, 2009b), rail covers most of the UK. Moreover, train and the car modes dominate travel in/out of Greater London with 790 and 320 thousand in-commuters and out-commuters daily in 2008. The 2.1 million daily trips on average carried out in only Greater London alone show that the train mode of transport is a very popular way of travel for long distance travel. This makes it a mandatory mode to be tested for using NM.

8.5.1 Network

The rail network includes London's Underground and Over-ground networks. The network extends from Scotland, south to South England and west to Wales. TfL provides the rail network only within London; however, other data repositories are available online providing free user-generated data for the whole network. OpenStreetMap (OSM) is a good example of these online free data repositories. OSM provides the whole British Rail network covering most of the UK. Nevertheless, as previously mentioned in Table 8.3, OSM does not provide each route as a unique polyline. This means that a trajectory cannot be tested to whether it is following a single route or if it were jumping from route to another, where in the latter's case it should not be matched to the train network and hence is not classified as train. This is a limitation of the dataset which is provided by OSM and limits the usability of the results obtained from the train NM process. The dataset is presented in Figure 8.17 below showing its extent within the UK. As can be noted from the map, we only use data for England for the purpose of this test.

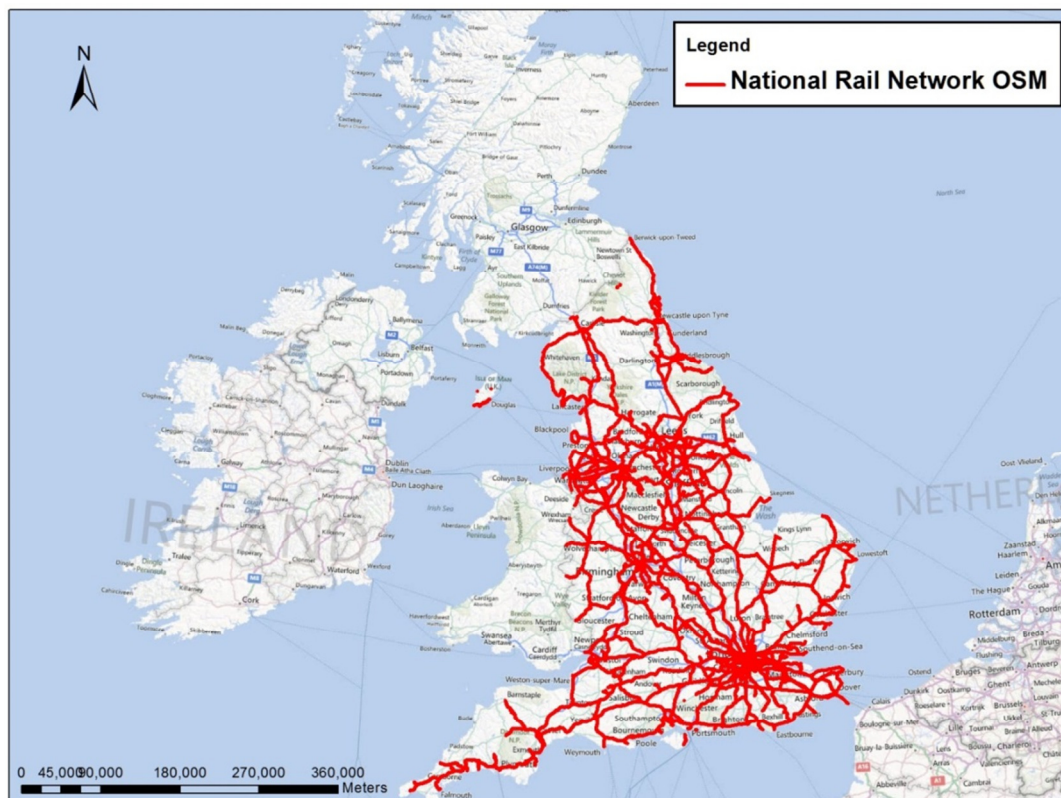


Figure 8.17 UK National Rail Network (Data provided by: OpenStreetMap)

8.5.2 Constraints

Similar to the underground NM process, there are two constraints for the train NM, namely; (a) the **maximum distance** allowed to the rail network and (b) the **minimum time** duration threshold that a given stage is allowed to be, for it to be qualified as a candidate for being matched to the rail network (as previously illustrated in Figure 8.10).

Similar to tube stages network matching; a 150 metre-distance threshold could be applied when testing for train trips. However, due to the absence of the rail network's topology information, the train mode would be more susceptible to error causing an over-classification where non-rail trajectories would be matched to the rail network due to their close proximity to rail tracks when they do not actually follow specific train routes. Therefore, one way this over-classification can only be reduced is by further restricting the distance error allowance. For that reason we use a lower distance threshold of 100 metres for matching to the rail network.

According to (TfL, 2013a), the average rail trip duration in 2011/2012 was found to be around 17% longer than that of London Underground for trips wholly within Greater London for entire trips where the main modes used are train and tube. These figures are only indicative of the scale of the difference between both modes since the figures reflect travel time of the whole trip for example including walks from these transportation hubs from and to origins and destinations. For the purpose of this study, rail trips will appear even longer than underground trips since we only account for underground trip sections that appear above the ground. Therefore, a minimum train stage time threshold of **5 minutes** was adopted as a value which is slightly higher than tube stages since we have no other means to calculate the minimum train trip travel time. This duration threshold is considered to be an assumption and a limitation of testing the algorithm on train trips.

8.5.3 Matching Accuracy

The rail NM algorithm produces more accurate results out of Greater London, as demonstrated in the examples shown in Figure 8.18 and Figure 8.19. This is due to the fact that London has a high density of rail tracks near to one another, and adding to that the fact that the rail network does not have a unique route identifier, the algorithm fails to match train trajectories to the rail network efficiently within London.

Figure 8.18 for example shows several trips made between London and Takeley where the track is matched very accurately since not many other train routes are nearby. Figure 8.19 too shows repetitive trips from London to North West England. The figure illustrates that there are other rail tracks nearby but not as dense as within London, which is shown in the south east of the map.

8 Phase III: Verification (Network Matching)

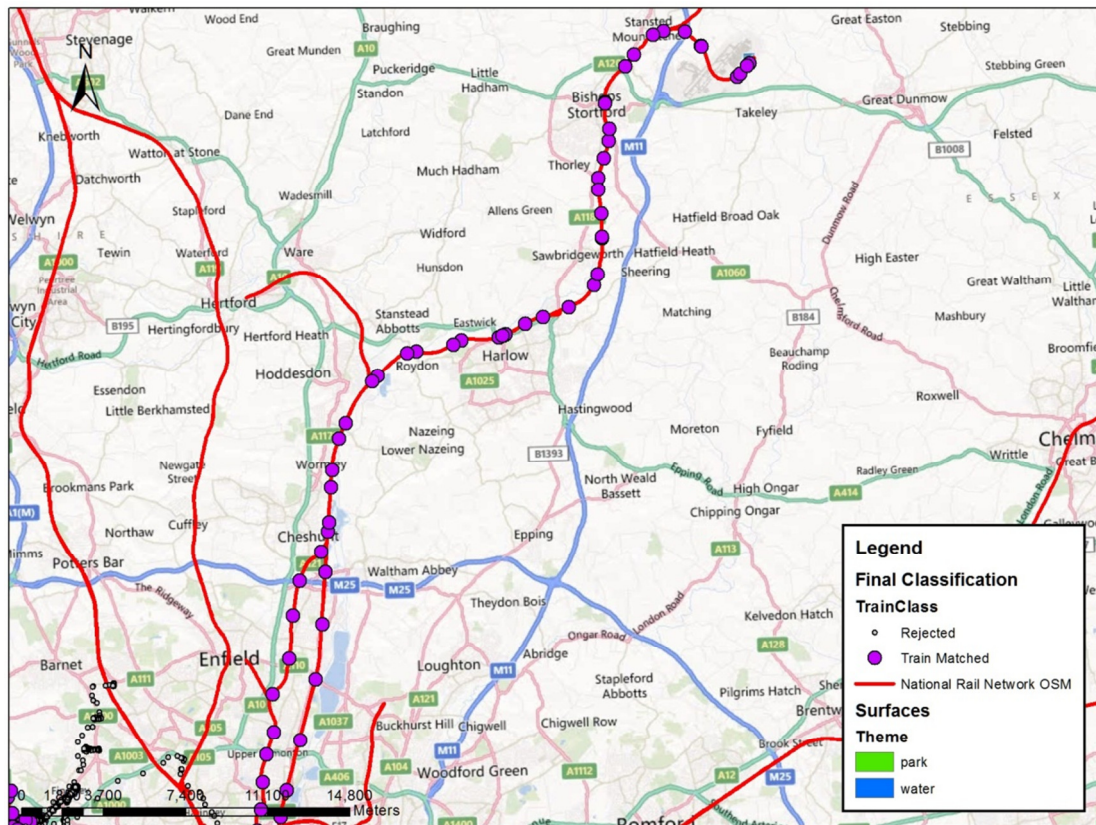


Figure 8.8 Example of Matching to UK National Rail Network

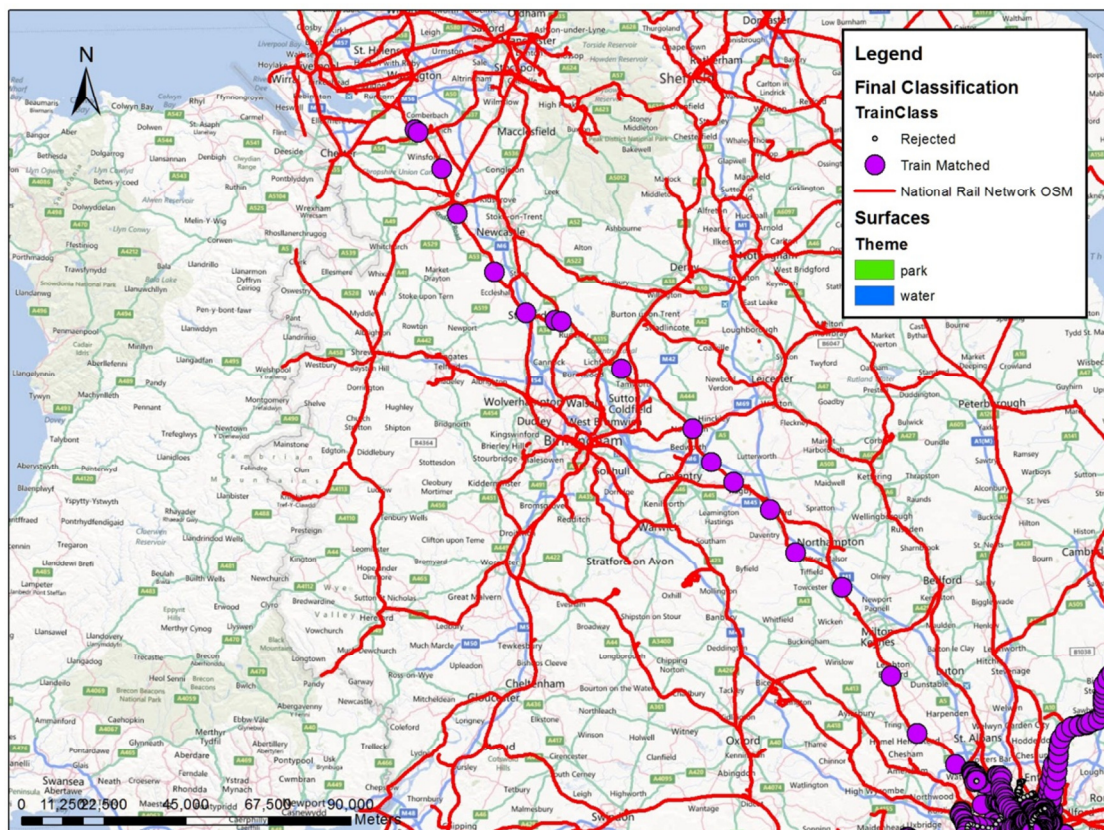


Figure 8.19 Another Example of Matching to UK National Rail Network

On the other hand, Figure 8.20 shows an example of a mismatch within London, due to the set distance threshold, although a trajectory is not moving on the rail network while still being close to a train route at every GPS fix yet it still is matched to the rail network. Such scenarios cannot be resolved since no route information is provided. Therefore, we use a decision-making phase at the end of the NM process to reason between the results attained from NM to different public transport networks. This is described in section 8.7 in full detail.

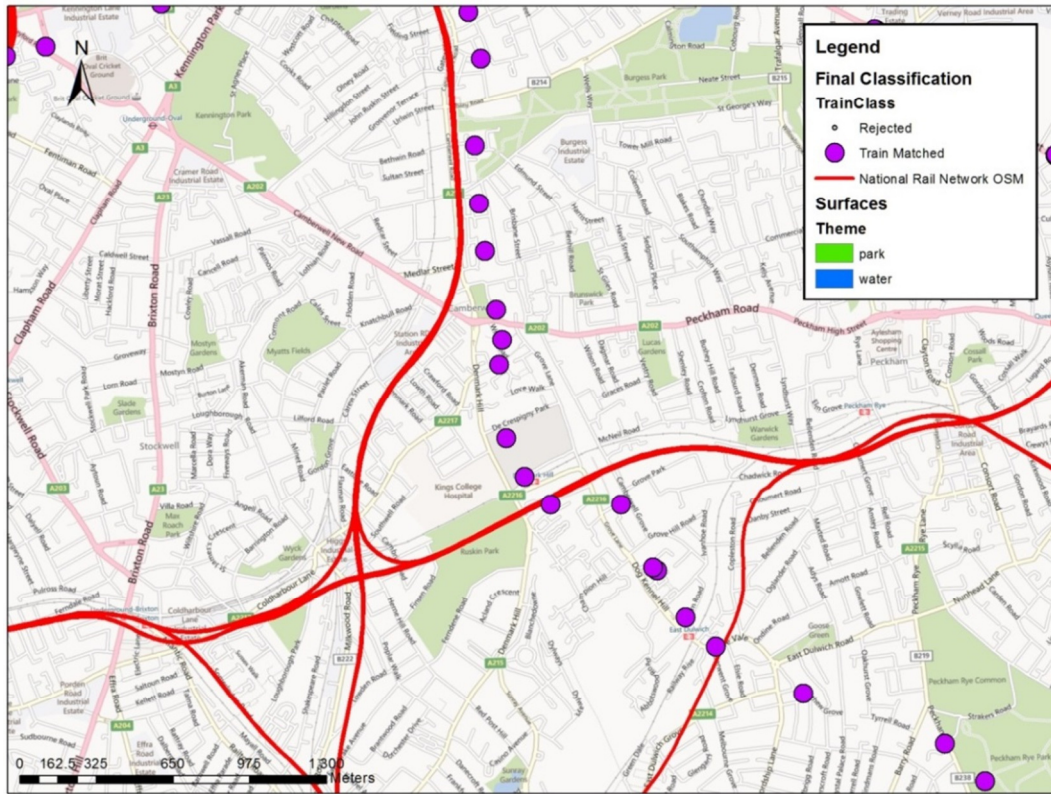


Figure 8.20 Example of Mismatching to UK National Rail Network due to absence of Route Information

Table 8.9 presents the confusion matrix for the results from the rail NM for GPS non-walk stages. It is worth to note that 68% of train stages were correctly identified using the rail NM algorithm. However, other modes were also matched to the rail network, although with much lower percentage, as a result of the absence of the routing information.

| Truth | Inferred | | | |
|-------|----------|-----------|-------|---------------------------|
| | No Entry | Not Train | Train | % of Train Matched Stages |
| bus | 5 | 124 | 25 | 16.23% |
| car | 12 | 111 | 87 | 41.43% |
| cycle | 0 | 63 | 27 | 30.00% |
| train | 1 | 7 | 17 | 68.00% |
| tube | 38 | 30 | 15 | 18.07% |
| walks | 579 | 0 | 0 | 0.00% |

Table 8.9 Rail Network Matching of GPS Stages Confusion Matrix Results

8 Phase III: Verification (Network Matching)

On the other hand, Table 8.10 presents the confusion matrix for the results from the rail NM for GPS non-walk stages resulting from SVM classification. The table shows that all the train stages were detected using the rail NM algorithm (100%), yet still some stages of other modes were also matched to the rail network.

| Truth | No Entry | Inferred | | |
|-------|----------|-----------|-------|---------------------------|
| | | Not Train | Train | % of Train Matched Stages |
| bus | 44 | 86 | 24 | 15.58% |
| car | 49 | 38 | 123 | 58.57% |
| cycle | 17 | 51 | 22 | 24.44% |
| train | 0 | 0 | 25 | 100.00% |
| tube | 58 | 17 | 8 | 9.64% |
| walks | 389 | 107 | 83 | 14.34% |

Table 8.10 Rail Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results

As a result of the rail NM phase, all the train stages were matched correctly, yet some of other modes were also matched incorrectly. This conflict will be resolved in a reasoning stage presented in section 8.7. Next section presents the NM to the bus network, discussing the network, algorithm constraints and results.

8.6 Bus Network Matching

Approximately 7,500 buses carry more than six million passengers each weekday on a network serving all parts of Greater London (TfL, 2012b). That makes the bus network one of the densest networks in London after the road network. This makes it essential to include such a network in the NM phase, where many bus trips can be differentiated and distinguished from other modes.

8.6.1 Network

More than 90% of Londoners live within 400 metres of one of the 19,500 bus stops in the Capital (TfL, 2012b). This implies that the bus network is well distributed across London. The 1464 bus routes that constitute the bus network in London are shown in Figure 8.21. The bus network data is provided by TfL for the purpose of this research.



Figure 8.21 London TfL Bus Network (Data provided by TfL)

8.6.2 Constraints

Similar to the underground and train NM processes, the two constraints for the bus NM are (a) the **maximum distance** allowed to the bus network to allow the inclusion of a bus route, and (b) the **minimum time** duration threshold.

In this case, we set the distance threshold similar to the underground NM process (**150 meters**), where the distance depends on the GPS device accuracy and GPS systematic error sources, which are similar for the three modes.

As for the temporal constraint, we adopt a similar time threshold for bus stages as previously used for train trips (**5 minutes**). Using a similar time threshold is meant to serve as means to test the performance of the network matching algorithm consistently for bus and train modes due to complexities in both their movements. For train movements, this complexity is a result of the absence of network topological information for train movements. Bus movements also face the difficulty of identifying route switch points within bus trips since bus stops are in an opened road environment where passengers are free to alight one bus route to walk to other bus stops and switch bus routes.

8.6.3 Matching Accuracy

Two examples of the results of the bus NM process are shown in Figure 8.22 and Figure 8.23. The first shows an individual's everyday habitual (repetitive) trip where the participant uses the bus frequently, noted from the density of the GPS points on the bus route. The latter figure shows some bus network-matched trips within central London (in blue), and all other non-matched trips (in grey) clearly use other networks such as the road network.

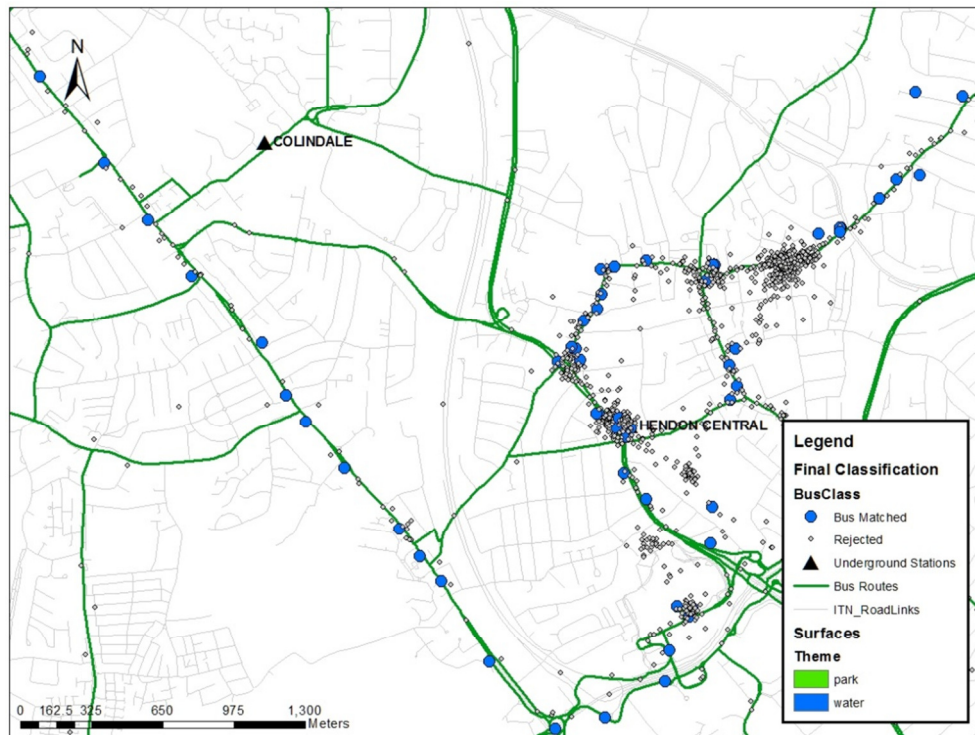


Figure 8.22 Example of Network-Matched GPS Fixes to London Bus Network

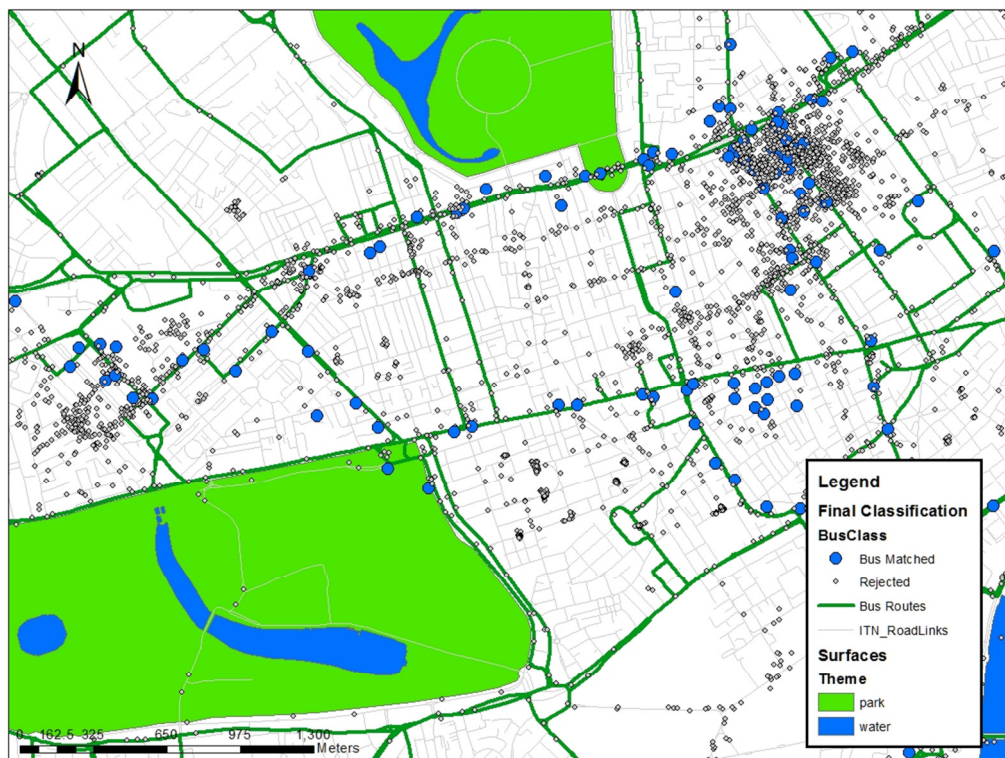


Figure 8.23 Another Example of Network-Matched GPS Fixes to London Bus Network

An example of the output file produced as a result of the bus NM algorithm is illustrated in Table 8.11. The CSV file produced contains a column containing the decision of whether the stage is matched to the bus network or not or not considered to enter the algorithm in the

first place. In case of a bus network match, seven columns are filled with the 7 nearest matched bus routes.

| ID | Bus Class | Bus Route1 | Bus Route2 | Bus Route3 | Bus Route4 | Bus Route5 | Bus Route6 | Bus Route7 |
|------|-----------|------------|------------|------------|------------|------------|------------|------------|
| 5236 | o | NA | NA | NA | NA | NA | NA | NA |
| 5237 | o | NA | NA | NA | NA | NA | NA | NA |
| 5238 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5239 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5240 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5241 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5242 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5243 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5244 | bus | 73 RUN2 | 24 RUN2 | 134 RUN1 | N253 RUN1 | N5 RUN1 | N279 RUN2 | N73 RUN2 |
| 5245 | o | NA | NA | NA | NA | NA | NA | NA |
| 5246 | o | NA | NA | NA | NA | NA | NA | NA |
| 5247 | o | NA | NA | NA | NA | NA | NA | NA |
| 5248 | o | NA | NA | NA | NA | NA | NA | NA |
| 5249 | not bus | NA | NA | NA | NA | NA | NA | NA |
| 5250 | not bus | NA | NA | NA | NA | NA | NA | NA |
| 5251 | not bus | NA | NA | NA | NA | NA | NA | NA |
| 5252 | not bus | NA | NA | NA | NA | NA | NA | NA |

Table 8.11 Example of the Output from the Bus Network Matching Process

Since the bus network coincides with the road network, a great amount of confusion is expected. The car and cycle modes are expected to dominate this confusion, since they run on the same network. This is evident from Table 8.12, where 16% of the car mode and 12% of the cycle mode stages are matched to the bus network. Nevertheless, 55% of the bus stages were matched correctly to the bus network.

| Truth | Inferred | | | |
|-------|----------|---------|-----|-------------------------|
| | No Entry | Not Bus | Bus | % of Bus Matched Stages |
| bus | 5 | 65 | 84 | 54.55% |
| car | 12 | 165 | 33 | 15.71% |
| cycle | 0 | 79 | 11 | 12.22% |
| train | 1 | 24 | 0 | 0.00% |
| tube | 38 | 39 | 6 | 7.23% |
| walks | 579 | 0 | 0 | 0.00% |

Table 8.12 Bus Network Matching of GPS Stages Confusion Matrix Results

On the other hand, Table 8.13 shows that only 22% of the bus stages classified using the SVM algorithm were matched to the bus network, while having a negligible amount of confusion. This makes the bus NM algorithm quite efficient and not over-classifying. Hence, if a non-walk stage were to be matched to the bus network, then it is probably a bus stage. This is useful in the reasoning stage to decide on the mode assignments from the NM results (described in section 8.7).

| Truth | Inferred | | | |
|-------|----------|---------|-----|-------------------------|
| | No Entry | Not Bus | Bus | % of Bus Matched Stages |
| bus | 46 | 74 | 34 | 22.08% |
| car | 52 | 157 | 1 | 0.48% |
| cycle | 17 | 71 | 2 | 2.22% |
| train | 1 | 24 | 0 | 0.00% |
| tube | 59 | 24 | 0 | 0.00% |
| walks | 415 | 143 | 21 | 3.63% |

Table 8.13 Bus Network Matching of GPS Stages Confusion Matrix applied to SVM Classification Results

8.7 Modal Classification Verification for Public Transit Network Matching

Once the NM for different public transport networks is applied, logical reasoning is applied to the results need and a decision on the final classification of a stage needs to be made. The decision in this case is made on the row-level. This means for example if a GPS segment is classified using SVM as bus and is matched to both the bus and the underground networks, a decision is made for this segment only from the results of the conducted tests. Once this is applied on every segment of the stage, the stage might then consist of car and underground segments. Hence, a second decision on the stage as a whole needs to be made to avoid multi-modal stages, which are evidently incorrect.

Therefore, in this final part of the algorithm we conduct two phases of adjustments, namely; **(1) a horizontal (or a segment-wise) adjustment**, which treats each segment as an entity on its own deriving its final classification from results of SVM classification and different NM tests. And **(2) a vertical (or a stage-wise) adjustment**, which treats each stage as an entity on the whole of its segments, and derives a further adjustment of the final classification of the whole stage from the individual classifications of its segments. The process also includes a re-segmentation phase in the middle to re-calculate the segments after their change due to applying the NM results.

8.7.1 Segment Modal Decision Making Phase: Segment-Wise Horizontal Reasoning

Since more than 90% of Londoners live within 400 metres of one of the 19,500 bus stops in the Capital (TfL, 2012b), the bus network is accounted as a very well distributed network across London. Moreover, TfL (2009b) states that the bus mode constituted 19% of the modal share of the daily journey stages in London in 2007, while the underground 10% and rail 8% (illustrated in Figure 8.24). This suggests that in the case of conflict between the public transport modes, the probability of a trip being of the bus mode is the highest followed by the underground then rail modes respectively. We use this fact to differentiate between the results of the NM of the different public transport modes.

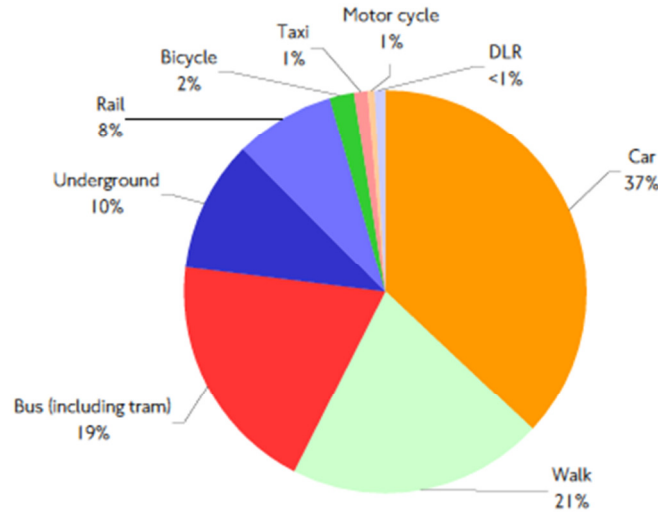


Figure 8.24 Modal Shares of Daily Journey Stages in London, 2007 (Source: (TfL, 2009b))

Table 8.14 shows an evident example of this conflict showing a case where a stage of a non-walk trajectory was matched to the 205 RUN₁ service on the bus network and to the Circle line on the underground network. Such scenario evolves due to the vicinity of certain bus and underground tracks to one another. This kind of information can be used for applications such as feedback transport systems to propose to commuters alternative means of transport for their habitual trips. Nevertheless, in this research this conflict in mode verification needs to be resolved using a logical reasoning procedural algorithm.

| ID | Bus Class | Train Class | Tube Class | Tube Route ₁ | Bus Route ₁ |
|------|-----------|-------------|------------|-------------------------|------------------------|
| 2552 | o | o | o | NA | NA |
| 2553 | o | o | o | NA | NA |
| 2554 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2555 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2556 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2557 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2558 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2559 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2560 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2561 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2562 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2563 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2564 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2565 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2566 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2567 | bus | not train | tube | Circle | 205 RUN ₁ |
| 2568 | bus | o | tube | Circle | 205 RUN ₁ |
| 2569 | o | o | o | NA | NA |
| 2570 | o | o | o | NA | NA |

Table 8.14 Example of Network Matching Results (matching to Bus & Underground Networks)

8 Phase III: Verification (Network Matching)

Therefore, a decision matrix is developed to deal with reasoning between the results of the NM processes of the different public transport networks, in order to agree on the final modal classification decision in different cases. Table 8.15 applies this ideology and also bases the decisions on the accuracy figures from NM and SVM classification of the pilot data previously discussed in sections of this chapter.

| Speed Classification | Acceleration Classification | Bus Network Matching | Tube Network Matching | Train Network Matching | Time-Distance Tunnel Travel | Final Classification | Rationale |
|------------------------------------|------------------------------------|----------------------------------|-----------------------|------------------------|-----------------------------|------------------------------------|---|
| | | | | | Tube | Tube | Due to the high accuracy of the Time-Distance algorithm |
| Bus Bus Bus X Bus X | Bus Bus X Bus X Bus | Bus - - - Bus Bus | | | | Bus Car X X Bus Bus | Stemming from the fact that Bus NM is of a high accuracy, otherwise other modes will be chosen, and if not, then the car mode is the closest. |
| Cycle | Car | | | | | Car | Acceleration is better for discriminating Car from Cycle modes |
| | | - | Tube | | | Tube | Because Tube NM is very accurate if not matched to Bus |
| Train X | X Train | - - | - - | - - | | X X | Usually Train NM is over-classified, and hence if not matched, then it is definitely not Train. |
| Train | Train | - | - | - | | Car | Car is the mode with the highest confusion with Train mode. |
| Train | Train | - | - | Train | | Train | The train mode comes last due to the absence of a unique identifier, and therefore it is the last option if there is more than one network chosen for the NM. |

X: indicates that another non-walk mode was classified other than the mode under investigation.

-: indicates that the network matching process did not verify the mode under investigation.

Table 8.15 Decision Matrix of Final Classification Assignment Based on Network Matching Results

The table is designed as a decision matrix from top to bottom in a procedural manner, where if the top row is not satisfied the next row is tested and so forth. For example, in the second row, if the SVM classifications based on speed and acceleration both return bus mode, and the bus NM returns a bus network match, then the final decision will be a bus classification regardless the rest of the NM results (tube and train). However, if this is not satisfied, the next row (next segment) is tested and so forth. The “Rationale” column in Table 8.15 provides an explanation why these decisions have been made.

8.7.2 Re-Segmentation Phase

The segmentation process segments the classification into separate stages, each of a unique mode of transport. Hence, the segmentation process is considered as a by-product of the classification process. The process is explicitly described in detail previously in chapter 7; however, it is re-used here to re-identify the recently amended stages due to the NM process.

This process groups consequent similar-mode segments and creates stages out of them as a result. The purpose of this process is to prepare newly-amended stages for the next phase, which is vertical adjustment. As will be described in the next subsection 8.7.3, the vertical adjustment will study the whole stage as one entity and apply adjustments based on reasoning of the entity as a whole.

The process first starts with separating the non-walk segments from the walk segments. It then groups non-walk segments creating different stages. Each non-walk segment can consist of several modes of transport. These conflicts will be resolved in the next phase of vertical adjustment (subsection 8.7.3). The algorithm then goes through every stage and notes the first segment in each stage to facilitate the search for stages, and hence, reducing the computation time.

8.7.3 Sequence Adjustment Phase: Stage-Wise Vertical Reasoning

After each GPS segment is reasoned for in the horizontal reasoning phase, the stage as a whole needs to be reasoned for. In this phase, each non-walk stage is analysed as a single entity. Several tests are conducted to amend a stage that might be a mix of segments of different modes. The following tests/amendments are applied to walk and non-walk stages:

1. If a long non-walk stage is mostly a certain mode (e.g. Underground) and instances of any other mode, it shall be replaced by the most occurring mode (i.e. Underground in this case) (Figure 8.25).

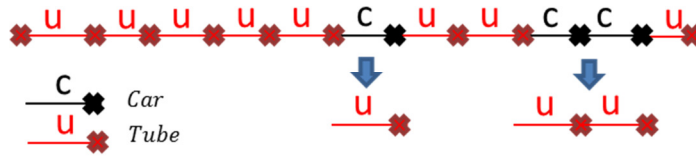


Figure 8.25 Amending Mixed Non-Walk Stages

2. Gaps between an Underground stage and an Underground instance (or for any other mode) shall be filled by Underground, as illustrated in Figure 8.26. This is not only restricted to the underground mode, but to all other non-walk modes. In other words, we merge any two very close stages of the same mode separated by maximum a gap of a threshold time and a threshold distance of a walk-stage. We have given the time constant a value of 3 minutes which is triple the epoch rate of collection for a maximum 3 walk segments that could be the time taken for a traffic light to change or the delay time of vehicles stopping can be as long 150 seconds in the evening congestion peaks (TfL, 2009a). We have assigned the distance threshold a value of 50 meters which is around the GPS accuracy that could be attained using either the used GTrek or u-blox devices (GTrek, 2012) (u-blox, 2009) plus GPS errors caused by urban canyons (Hinch, 2007).

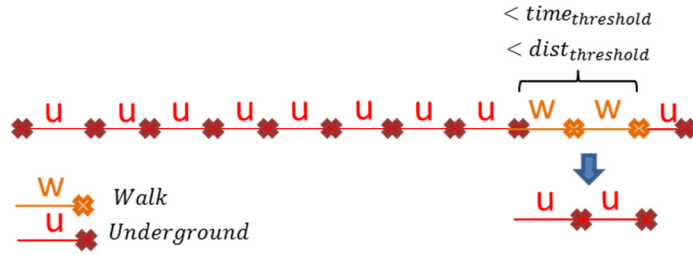


Figure 8.26 Merging Similar Non-Walk Stages Separated by Short durations of Walk Stages

3. Since the SVM classification stage classifies threads of three GPS segments, there are instances where some non-walk stages are classified correctly but contain a couple of over-classified segments at the beginning or at the end of the stage. These over-classified segments, if at the beginning of the non-walk stage, are originally the end of the previous walk stage, and if at the end, are originally the beginning of the next walk stage. This is resolved by identifying these over-classified segments and re-classifying them as walk segments. Figure 8.27 below presents an example of a public transport stage (Underground), demonstrating that if speed of the first two or last two segments do not exceed the assigned walk speed threshold, then these segments are re-classified as walk segments.

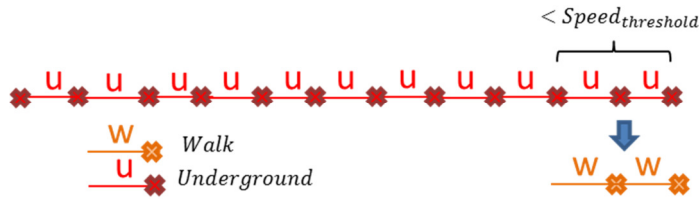


Figure 8.27 Dismissing Extra Over-Classified Walk Segments from Ends of Non-Walk Stages

The reason why the algorithm considers maximum only two segments for testing is because the SVM classification was applied to every three consecutive segments together, and a misclassification can only occur if one or two segments were walk segments. On the other hand, if the three segments were classified as walk, then the thread would have probably been classified as walk mode since walk segments are of a distinctively different speed and acceleration ranges. The upper speed threshold for rectifying segments back to walk mode is set to be 1.6 m/s based on figures from a report studying Pedestrian Level of Service design and impact on the quality of pedestrian life (Bloomberg & Burden, 2006). Expressing these figures, Weidmann (1993) obtained a normally distributed average speed of 1.34 m/s and a standard deviation of 0.26 (CV = 19%) for pedestrians walking on the street.

4. If a segment within a non-walk stage is classified by the underground travel time-distance detection method, the rest of the stage is classified accordingly to underground, as illustrated in Figure 8.28. This is due to the fact that the time-distance underground method is more accurate than results from SVM classification, and therefore, the classification is amended to underground mode.

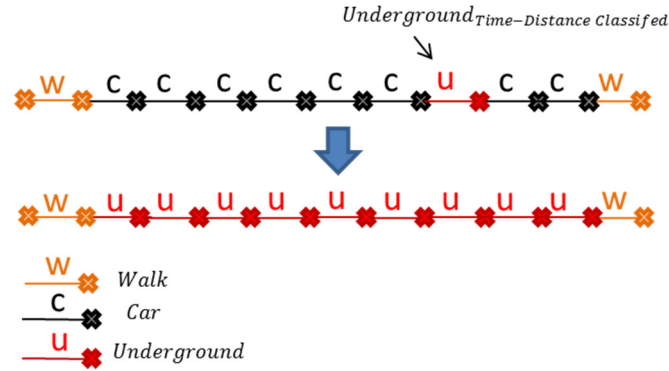


Figure 8.28 Amending Non-Walk Stages that contain a Segment Classified as Underground Mode using the Time-Distance Algorithm all into Underground

8.8 Network Matching Accuracy

After all the adjustments are applied to the SVM classification results, the final classification is assigned to each stage. Table 8.16 shows the confusion matrix of the results from the SVM classification phase, earlier presented in chapter 6. On the other hand, Table 8.17 shows the same confusion matrix but after applying the NM verification phase. It can be noted that there is an increase in the accuracy of most modes. The highest accuracies achieved are those of tube and cycle modes. The achieved accuracies before and after the NM process are 68% and 73% respectively. The increase of the overall accuracy is around 5%. Excluding the walk mode as an ancillary mode, the increase in accuracy is around 12% after the NM process.

| | | Actual | | | | | | Total Count |
|----------------|-------|--------|--------|--------|--------|--------|--------|-------------|
| | | bus | car | cycle | train | tube | walk | |
| Classification | bus | 38.60% | 18.56% | 16.92% | 12.50% | 3.96% | 7.52% | 112 |
| | car | 27.19% | 35.05% | 3.08% | 50.00% | 22.77% | 5.31% | 118 |
| | cycle | 21.05% | 25.77% | 72.31% | 0.00% | 4.95% | 3.10% | 115 |
| | train | 0.00% | 7.22% | 0.00% | 37.50% | 0.00% | 0.00% | 10 |
| | tube | 3.51% | 2.06% | 0.00% | 0.00% | 64.36% | 0.44% | 73 |
| | walk | 9.65% | 11.34% | 7.69% | 0.00% | 3.96% | 83.63% | 409 |
| Total Count | | 114 | 97 | 65 | 8 | 101 | 452 | 837 |

Table 8.16 Confusion Matrix of SVM Results Before Network Matching

| | | Actual | | | | | | Total Count |
|----------------|-------|--------|--------|--------|--------|--------|--------|-------------|
| | | bus | car | cycle | train | tube | walk | |
| Classification | bus | 46.96% | 13.73% | 10.14% | 7.14% | 0.00% | 9.61% | 120 |
| | car | 20.87% | 47.06% | 7.25% | 42.86% | 4.88% | 5.46% | 112 |
| | cycle | 13.91% | 22.55% | 73.91% | 14.29% | 1.22% | 2.18% | 103 |
| | train | 0.00% | 3.92% | 0.00% | 35.71% | 0.00% | 0.00% | 9 |
| | tube | 8.70% | 1.96% | 1.45% | 0.00% | 89.02% | 0.44% | 88 |
| | walk | 9.57% | 10.78% | 7.25% | 0.00% | 4.88% | 82.31% | 408 |
| Total Count | | 115 | 102 | 69 | 14 | 82 | 458 | 840 |

Table 8.17 Confusion Matrix of SVM Results After Network Matching

Table 8.18 presents the confusion matrix as the difference between Table 8.17 and Table 8.16, to illustrate the changes in accuracy of classifications and misclassifications before and after the NM process. As for the accuracy of classification, the table reveals an increase in all modes except the train mode. This is due to the increase of the number of the train stages from 8 to 14 trips (nearly twice), as well as due to the lack of unique train route identifier. On the other hand, there is a huge rise in accuracy in identifying tube, bus and car modes due to the NM process (25%, 8% and 12% increase respectively).

| | | Actual | | | | | |
|---|-------|--------|---------|--------|---------|---------|--------|
| | | bus | car | cycle | train | tube | walk |
| Classification | bus | -8.36% | 4.83% | 6.78% | 5.36% | 3.96% | -2.08% |
| | car | 6.32% | -12.01% | -4.17% | 7.14% | 17.89% | -0.15% |
| | cycle | 7.14% | 3.22% | -1.61% | -14.29% | 3.73% | 0.91% |
| | train | 0.00% | 3.29% | 0.00% | 1.79% | 0.00% | 0.00% |
| | tube | -5.19% | 0.10% | -1.45% | 0.00% | -24.67% | 0.01% |
| | walk | 0.08% | 0.56% | 0.45% | 0.00% | -0.92% | 1.31% |
| Total Count | | 115 | 102 | 69 | 14 | 82 | 458 |
| Notes: Difference in Accuracy of Mode Classification (Diagonal): Negative Values: Excellence of SVM + NM Positive Values: Excellence of SVM Only Difference in Accuracy of Mode Misclassification (the rest): Blue: SVM + NM Reduces Misclassification (positive values) Red: SVM only Reduces Misclassification (negative values) | | | | | | | |

Table 8.18 Before & after Network Matching Accuracy Difference expressed in Stages

Table 8.18 also shows that most confusions/misclassifications have been reduced (in blue). There is only an increase in over-classifying cycle into train trips (in red). In order to assess the change in accuracy appropriately, the classification and misclassifications of each mode need to be analysed aggregately. This is achieved by calculating the difference between Type I and Type II errors of the results of before and after NM. A Type I error is the rejection of a potentially true null hypothesis (Sheskin, 2004). In this case, it is the inverse of the accuracy level of each mode. Therefore, to calculate the difference between Type I errors of before and after the NM process, we assign the difference in accuracy between both states presented in the first column in Table 8.19. A Type II error (second column in Table 8.19) is the same but for the sum of misclassifications of each mode, since it represents the failure to reject a false null hypothesis (Sheskin, 2004). What could be noted from the table is that aggregating the differences between the Type I and II errors together is that nearly all modes achieve a better overall classification. This even applies to the train mode that does not achieve high classification accuracy, yet achieves a decreased misclassification as a result of the NM.

| Mode | Type I Error Difference | Type II Error Difference | Type I & II Errors Difference |
|----------------|-------------------------------|--------------------------------|-------------------------------------|
| bus | 8.36% | 18.84% | 27.20% |
| car | 12.01% | 27.04% | 39.05% |
| cycle | 1.61% | 0.72% | 2.33% |
| train | -1.79% | 3.29% | 1.51% |
| tube | 24.67% | -6.53% | 18.14% |
| walk | -1.31% | 0.17% | -1.15% |
| Average | 7.26 % | 7.26 % | 14.51 % |

Blue (+ve) values demonstrate NM's excellence

Red (-ve) values demonstrate before NM's excellence

Table 8.19 Type I & Type II Errors Difference of before & after Network Matching

8.9 Chapter Summary

In this chapter we introduce the third phase in the mode of transport detection framework, namely, network matching. The network matching stage acts as a verification stage for the public transport modes that were detected by the initial classification in the first two phases of the inference framework. The network matching phase is applied to existent transport networks in London which have routing properties (identified routes) such as the underground, bus and train networks.

The matching is applied using a unified algorithm for all three networks, only varying certain threshold values in terms of the displacement distance of GPS fixes from any given network and the time duration of the whole tested stage of GPS fixes. However, the underground network is partially tunnelled where there is no GPS coverage. Hence, the network matching algorithm uses the locations of the first and last fixes before entering into a tunnel part of the underground network to match them to the nearest underground stations and make an assessment whether this part of the GPS track was on the underground network or not.

The accuracy of the mode of transport classification after applying the network matching process as a verification stage increases by 7.26% when tested on the pilot data. Figures show that a significant increase in accuracy occurs in identifying tube, bus and car modes due to the NM process (25%, 8% and 12% increase respectively). The train network matching process is not very efficient in London's case due to the absence of routing information which leads to achieving low accuracy.

By the knowledge of information such as driving license, ownership of a car, ownership of a bike and access to Barclays Bikes, the car and cycle conflict can be resolved with a greater accuracy. Moreover, classification of other modes will be enhanced too with other types of information like feedback from the users, such as for example; a user did not use the underground service all through the testing period.

Chapter 9 discusses the application of the framework developed in this research to the validation dataset of (95 participants), and uses the apriority knowledge of accessibility to different travel modes to understand the modal diversity of the dataset population.

Chapter 9

Further Validation (Results & Limitations)

9 FURTHER VALIDATION (RESULTS & LIMITATIONS)

Chapters 6, 7 and 8 described the inference framework and tested each one of its phases using the pilot dataset collected in the beginning of this research. In this chapter, we test the framework on the bigger dataset that was collected according to the specifications described in chapters 4 and 5. We start this chapter by describing the sample characteristics of the collected validation dataset. We then use this dataset to assess the accuracy of the inference framework in the order of consequent phases of the framework accumulatively. Finally, we discuss some issues highlighted by the inference results, and potential extensions of our study to account for these situations.

9.1 Validation Dataset

This section describes the GPS dataset collected for validating the inference framework developed in this thesis in terms of its adherence to specifications we set in chapter 4 on sample composition. It then analyses the sample characteristics in relevance to its socio-demographical information, device handling, amount of data per participant, and levels of accessibility to travel means along with its effect on the sample formation. This section also discusses movement and the average speeds obtained and the geographical data extent and scope of this study.

9.1.1 Data Specifications

As previously mentioned, the pilot data consists of 21 participants over 2 weeks using u-blox GPS devices (u-blox, 2009) at a collection rate of 1 minute, and the mode of transport was manually labelled by the participants segment-by-segment. However in chapter 4, we provided an understanding of the need of obtaining a sample that is well distributed across the study area in different times of the day and we have demonstrated how the validation dataset collected for this research consisting of 95 participants for a period of more than 2 weeks is required satisfies this need. In this chapter, we further investigate the collected dataset which was collected using GTrek GPS devices (GTrek, 2012). Supplementary information such as socio-demographics was also collected. Modal information such as ownership of a bike, access to Barclays Bikes, driving license and ownership of a car was also collected to ensure maintaining a high level of information quality. Moreover, in order to ensure good GPS data quality, we have set the GPS devices at a collection rate of 30 seconds to be resampled into 1 minute later when being pre-processed.

Participants have been recruited from the population of London over 18 years old using a random sampling (stratified) recruitment process run from the UCL network and passed on outside of it. We were particularly interested in recruiting participants who work long hours or do shift work and those who live in households with children. Upon agreeing to take part in the study, the participant was asked to meet with the researcher to explain the details of the survey and the GPS recording device they would be asked to carry during the survey period works. The participants were also asked at this meeting to provide their personal information and details of locations they frequently visit and travel habits they have. The participants were also made aware that they can withdraw their data from the project at any time up until it is transcribed for use in the final report.

9.1.2 Data Profile

As a result of the data collection, 6,599,286 GPS records were collected for 84 people, as 11 participants dropped out in the process of collection for different personal reasons over the period from 24/May/2011 to 16/Mar/2012. As shown in Figure 9.1, the dataset contained a nearly equal gender split across all age groups. The figure also shows that the largest group is from the first three younger age groups, whereas, the dataset contained only 10 participants from the older age group. The 11 drop-outs of the experiment equally stem from different age groups.

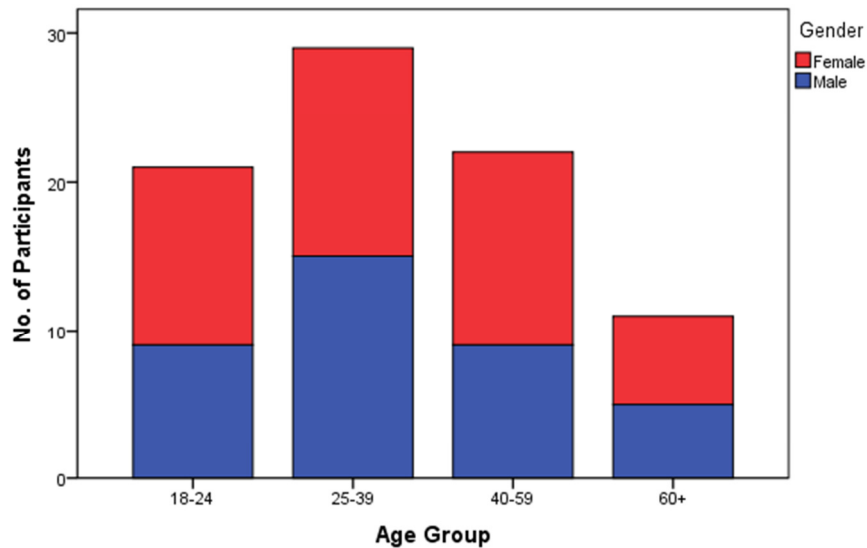


Figure 9.1 Gender Split across different Age Groups

Figure 9.2 shows the distribution of occupation nature types of the participants in this study separated by gender type. The figure also shows that as we initially planned for the study, the majority of the participants were out of the “full-time employed” and “full-time education” categories to ensure richness of the data. The figure also shows that within most categories there is a near-equal gender split, specifically in the two main categories. On the other hand, the “not working” and “retired” population is under-represented, specifically for the male gender. This does not impose a threat on the study as these categories would probably tend to have less movement and an assumption is made in this study that their absence would not significantly bias the obtained results.

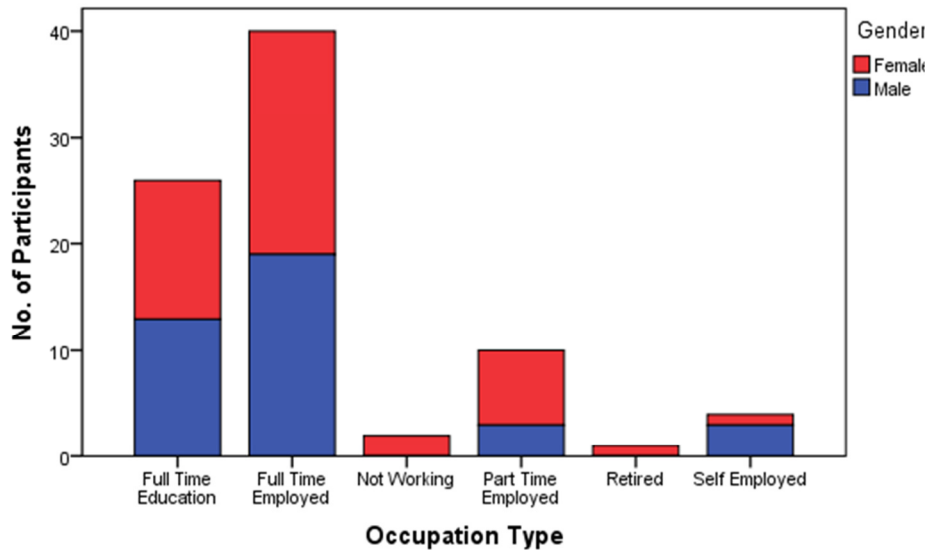


Figure 9.2 Occupation Nature Type of the Validation Dataset by Gender

9.1.3 Device Carrying Information

The amount of collected data depends on the travel patterns of each participant as well as on whether they choose to carry the device at all times. The amount of data collected by each participant does not affect the mode detection process since the framework does not use repetitive patterns of travel for classification. However, not using the device significantly would bear a shortage of the intended sample size required for testing the inference performance. Figure 9.3 shows a box plot of the device carrying durations per age group. The figure shows a larger variance in younger age groups which reflects the higher difficulty of retrieving the devices or ending the study earlier when dealing with younger groups.

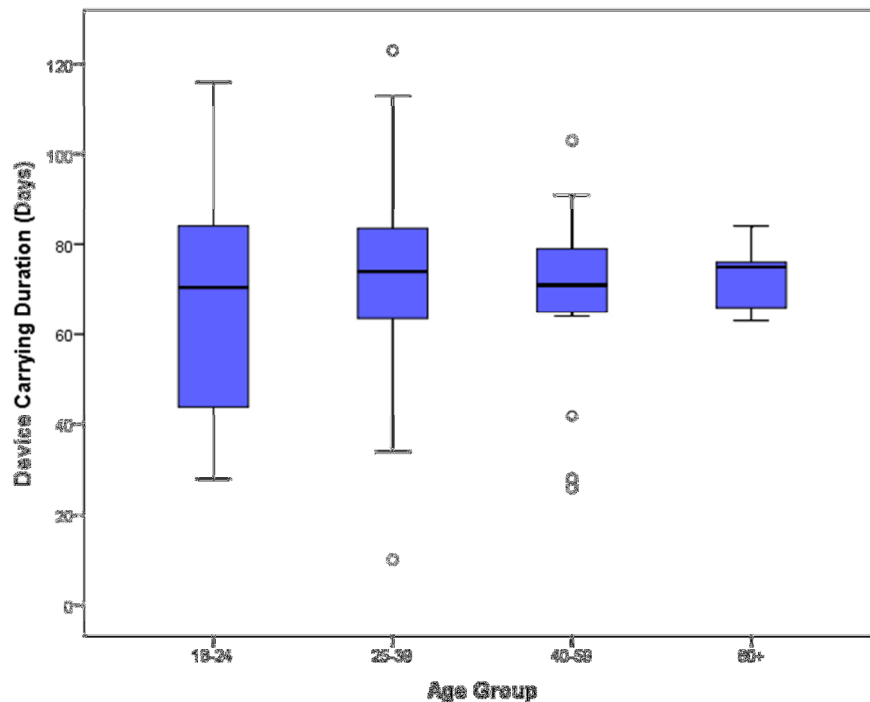


Figure 9.3 Device Carrying Time for different Age Groups

On the other hand, Figure 9.4 shows the number of GPS segments (or fixes) collected by the study's participants also divided by age group. The figure shows that older age groups tend to collect more GPS data. This might be due to that older participants are either more careful to carry their GPS devices in every trip they do or that they simply have more outdoor activity in their average day.

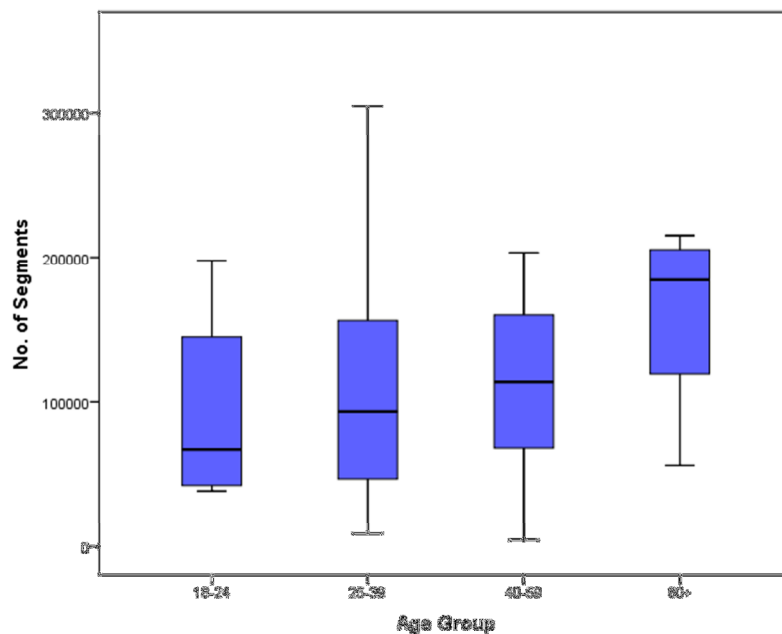


Figure 9.4 Number of GPS Segments/Fixes Collected by Age Group

9.1.4 Access to transport means

As previously mentioned, we have collected some modal-related information such as the access to different transport means. Figure 9.5 shows the ownership of participants from different age groups of elements that provide access to different transport means such as ownership of a driving license, car, bike, or membership of London's TfL Barclays Bikes cycling scheme. We can note from the figure that the majority of participants above 24 years old have a driving license and own a bike. The access to Barclays Bikes scheme appears to be minimal across all age groups. The ownership of a car, however, appears to be relatively popular among the 25-39 and 40-59 age groups. This mix of levels of access to transport means is beneficial in order to get a fair spread of usage of different modes of transport in this study.

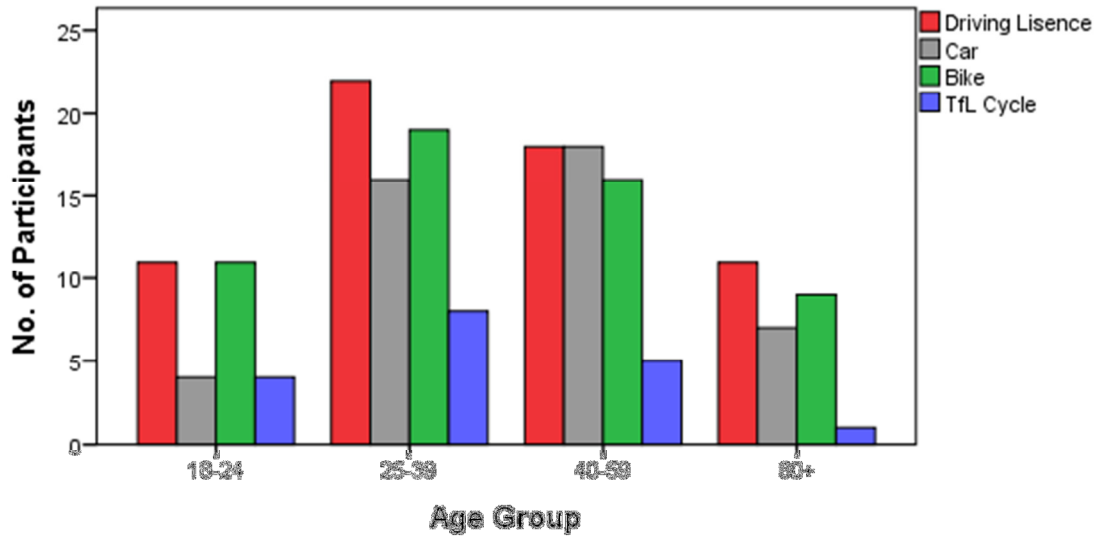


Figure 9.5 Participant Ownership of Access to different Transport Means by Age Group

9.1.5 Movement Statistics

A detailed description of the spatial, modal and temporal distributions of the dataset across different parts of London is described in chapter 4 (section 4.4) in order to understand the extent and the adequacy of the sample. In this section however we review other movement elements of the participants such as distances and speeds. Figure 9.6 plots the average length of GPS segments (distance between GPS fixes) versus the average time difference between GPS fixes. We might notice that the data of one specific participant contains several large gaps during data collection yet not travelling large distances, which is an indicator of that participant not using the device frequently or frequently being at indoor locations with signal blockage. The figure also reveals that around four participants have large average travelling distances between GPS fixes which might be due to the usage of air travel mode. The plot reveals that there is no huge sparseness in the temporal domain except for that one participant.

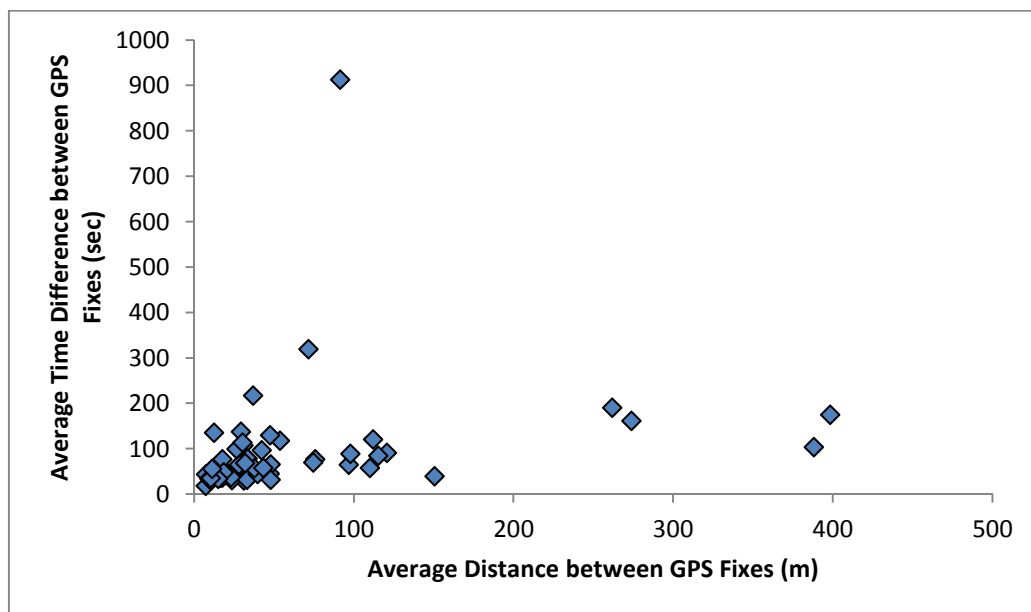


Figure 9.6 Average Distance vs. Time between GPS Fixes for each Participant

Figure 9.7 plots a histogram of the frequency of average speeds among the participant population. The histogram shows that most average speeds fall within 0 to 3 m/s which might reflect the domination of stops and walk modes. There are several other instances of average higher speeds that reflect the domination of faster modes in the activity of around 6 participants.

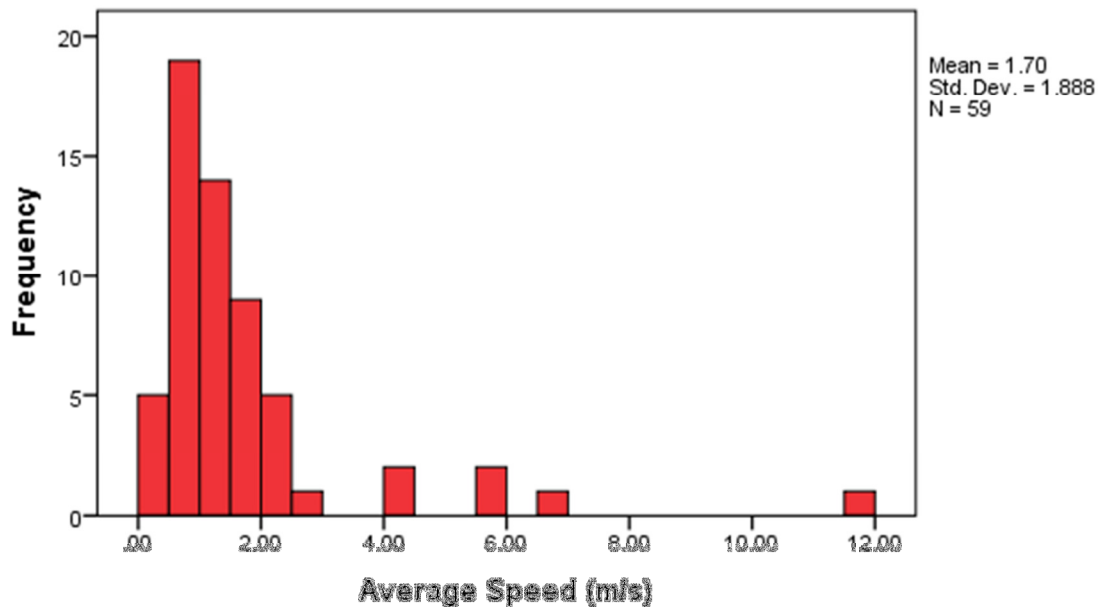


Figure 9.7 Histogram of Frequency of Average Speeds among Participants

9.1.6 Data Extent and Scope

All participants are mainly based in the city of London, however many of them reported several national and international trips within the study period. Figure 9.8 shows the geographic extent of the study which covers many parts of the UK and several international trips. As might also be noted, air trips are sometimes recorded entirely and in several occasions no flight information was recorded as seen in trips to Italy, Scotland, Denmark, etc. This research does not count air travel as one of the modes of transport to be detected using the inference framework and therefore, would be excluded from the validation results. Nevertheless, air travel is not considered as a difficult mode to separate due to its distinct high speed values, and therefore, it is considered out of the scope of the study.

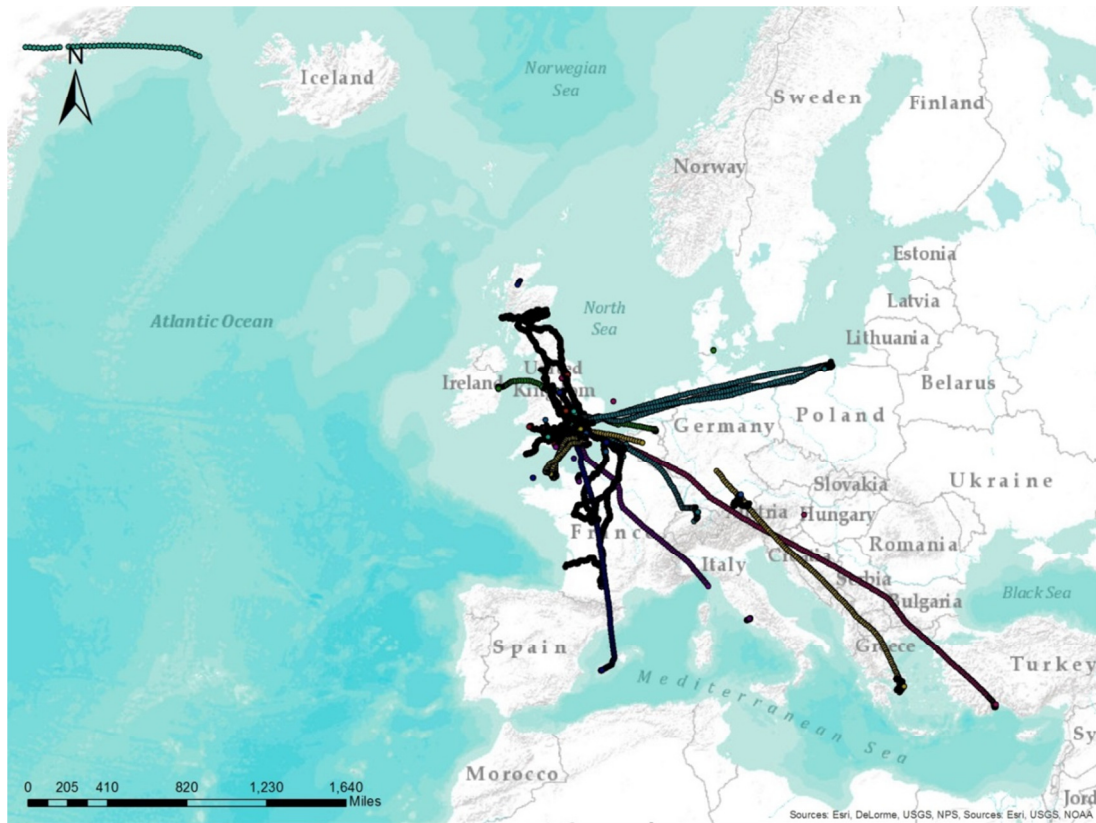


Figure 9.8 Geographic Extent of most of the Collected Data

It is also worth to mention that outside of London there would not be network matching to bus and underground transport networks because they are not available to this study. Therefore, bus and underground modes outside of London are also discarded from the validation process since they are out of the scope of the study. The overground mode in London was also excluded because its network dataset was provided later after the full processing of the data using the inference framework.

9.1.7 Data Summary

This section has described details of the specifications of the dataset collected in this chapter for validating the inference framework developed in this thesis. These specifications include providing a satisfactory coverage of the sample area in the spatial and temporal domains, adequate sample duration and collection rate specifications as set in chapter 4. We also described the composition of the sample dataset in terms of their socio-demographical information. We then described the distribution of device handling and movement characteristics in relation to these socio-demographics. Afterwards, we described accessibility characteristics to different modes of transport and the effect of levels of accessibility on the study's quality. We also briefly touched on analysing the distribution of movement and speeds across all participants. We finally then discussed the data extent and scope of the study relative to the geographical coverage of the collected GPS tracks.

9.2 Validation

This section describes the results of using the validation dataset collected in this chapter to assess the accuracy of different phases of the inference framework. The dataset is first split into two thirds for learning and one third for calculating the accuracy. The two thirds used for learning are then used for cross-validation to generate multiple training/validation set pairs (Alpaydin, 2004). The section will start by assessing the accuracy as result of applying the moving window SVM classification algorithm developed in chapter 6 and discussing the resulting confusions. This is followed by assessing the accuracy of stops detection after applying the spatio-temporal clustering algorithm developed in chapter 7. Finally, we present the results of applying the network matching algorithm developed in chapter 8, and we discuss the repercussions of applying network matching on the achieved detection accuracy.

9.2.1 Moving Window SVM-Based Classification Results

As described and tested in chapter 6, we have applied the moving window SVM-based classification on the pilot data yielding decent inference results. The classification was applied once for speed values and once for acceleration. We have then illustrated from the results that each variable has proven to be suitable for detecting certain modes better than the other. As a result, we applied a selection matrix that combines the results from both variables each identifying a set of modes. Here we apply the same classification algorithm to the validation dataset described in section 9.1.

We first use speed to apply the classification algorithm developed in this thesis achieving 98.19% general accuracy in detecting modes of transport which are bus, car, cycle, train, underground (tube), and walk. Moreover, if we focus on the non-ancillary modes by excluding the walk mode from the calculation, we calculate 70.72% general accuracy for the rest of the modes. Table 9.1 presents the confusion index of this classification. The table highlights the detection accuracy of each mode along the table's diagonal and the Type I and Type II classification errors in the rest of the table. Each cell in the table is also colour coded according to the value of accuracy or error it represents. As might be noted, the cycle, train and walk accuracies are relatively high compared to the rest of the modes. This can be attributed to the low speed values of walk and cycle modes and the high values of the train mode. There is some confusion noted between the bus and car modes due to the speed similarities within London. Many car trips were also confused as train mode due to speed similarities outside of London.

| | | Classification | | | | | | Total Count |
|-------------|-------|----------------|--------|--------|--------|--------|---------|-------------|
| | | bus | car | cycle | train | tube | walks | |
| Actual | bus | 78.16% | 4.13% | 8.53% | 0.85% | 8.08% | 0.24% | 22500 |
| | car | 21.67% | 47.92% | 1.22% | 24.05% | 4.46% | 0.69% | 62472 |
| | cycle | 13.37% | 1.02% | 85.21% | 0.00% | 0.11% | 0.29% | 43650 |
| | train | 0.00% | 1.04% | 0.00% | 96.38% | 2.04% | 0.53% | 23556 |
| | tube | 16.34% | 3.46% | 1.70% | 5.16% | 53.90% | 19.45% | 10242 |
| | walks | 0.19% | 0.04% | 0.04% | 0.07% | 0.03% | 99.63% | 3236004 |
| Total Count | | 44826 | 33330 | 41244 | 40782 | 11466 | 3226776 | 3398424 |

Table 9.1 Classification Results from applying the Moving Window SVM-based Classification Algorithm to Speed Values

On the other hand, when basing the classification on acceleration, the general achieved accuracy is 98.05% and 70.68% for non-ancillary modes, which are very similar to speed accuracies (see Table 9.2). The results in general look similar; however, modes like tube, train and bus seem to be better identified using acceleration. However, the car mode's detection accuracy appears to slightly drop based on acceleration classification, contrary to the results presented in Chapter 6. This might be due to that the pilot data did not have as many driving trips out of London as in the validation dataset. The reason of the fail of acceleration to identify trips outside of London as well as speed might be due to lack of changing speeds (in other words accelerating) when driving on motorways. On the other hand, drivers would tend to accelerate and decelerate a lot inside of London due to the constant congestion variation from one zone to the other.

| | | Classification | | | | | | Total Count |
|-------------|-------|----------------|--------|--------|--------|--------|---------|-------------|
| | | bus | car | cycle | train | tube | walks | |
| Actual | bus | 77.09% | 4.13% | 8.53% | 0.85% | 9.15% | 0.24% | 22500 |
| | car | 24.23% | 42.59% | 1.52% | 24.35% | 5.55% | 1.77% | 50178 |
| | cycle | 13.37% | 1.02% | 85.21% | 0.00% | 0.11% | 0.29% | 43650 |
| | train | 0.00% | 0.00% | 0.00% | 97.51% | 2.24% | 0.25% | 21432 |
| | tube | 17.35% | 2.55% | 1.80% | 4.85% | 55.85% | 17.60% | 9648 |
| | walks | 0.23% | 0.03% | 0.04% | 0.08% | 0.03% | 99.60% | 2743164 |
| Total Count | | 43206 | 23742 | 41244 | 35850 | 11574 | 2734956 | 2890572 |

Table 9.2 Classification Results from applying the Moving Window SVM-based Classification Algorithm to Acceleration Values

As a result of the supremacy of each variable in detecting certain modes, we apply a similar selection matrix to combine classification results from speed and acceleration to achieve an overall accuracy of 98.24%, and accuracy of 71.05% for non-ancillary modes. Table 9.3 shows the resulting confusion index of the combined results. It is clear that most modes are somewhat confused with bus and tube modes. Moreover, almost a quarter of car instances are confused as train due to the similarity in speed/acceleration values. These confusions can be resolved using network matching since the bus, tube and train modes follow their respective networks.

| | | Classification | | | | | | Total Count |
|-------------|-------|----------------|--------|--------|--------|--------|---------|-------------|
| | | bus | car | cycle | train | tube | walks | |
| Actual | bus | 78.80% | 4.11% | 8.53% | 0.85% | 7.57% | 0.13% | 22500 |
| | car | 21.60% | 47.89% | 1.22% | 24.00% | 4.62% | 0.67% | 62472 |
| | cycle | 13.35% | 1.02% | 85.14% | 0.00% | 0.21% | 0.29% | 43650 |
| | train | 0.00% | 1.04% | 0.00% | 96.15% | 2.37% | 0.43% | 23556 |
| | tube | 15.58% | 3.40% | 1.52% | 5.04% | 68.60% | 5.86% | 10242 |
| | walks | 0.19% | 0.04% | 0.04% | 0.07% | 0.03% | 99.63% | 3236004 |
| Total Count | | 44838 | 33300 | 41196 | 40680 | 13104 | 3225306 | 3398424 |

Table 9.3 Integrating Classification Results from applying the Moving Window SVM-based Classification Algorithm to Speed & Acceleration Values

9.2.2 Segmentation/Spatio-Temporal Clustering (Stops Detection) Results

The second phase of the inference framework aims to segment the track by mainly identifying stops using a spatio-temporal clustering algorithm we have developed. Long stays are also detected using a k-means clustering algorithm. Identified stops and long stays are then merged into the classification results in the inference framework. This is combined with the results from the classification phase to include a stationary mode added to the six existing modes this research aims to detect. Table 9.4 shows the confusion index as a result as combining the results from the classification phase and the segmentation. The table reveals an accuracy of 99.99% in identifying stops within the track. The overall accuracy of identifying the seven modes was found to be 98.35%, and 70.84% for non-ancillary modes.

| | | Classification | | | | | | | Total Count |
|-------------|------------|----------------|--------|--------|------------|--------|--------|--------|-------------|
| | | bus | car | cycle | stationary | train | tube | walk | |
| Actual | bus | 76.85% | 3.99% | 8.18% | 0.55% | 0.84% | 9.45% | 0.13% | 22728 |
| | car | 21.39% | 46.13% | 0.79% | 2.75% | 24.08% | 4.62% | 0.24% | 62244 |
| | cycle | 13.26% | 1.00% | 85.04% | 0.30% | 0.00% | 0.21% | 0.18% | 43650 |
| | stationary | 0.00% | 0.00% | 0.00% | 99.99% | 0.00% | 0.00% | 0.00% | 2979546 |
| | train | 0.00% | 0.99% | 0.00% | 0.41% | 96.10% | 2.37% | 0.13% | 23556 |
| | tube | 15.58% | 3.40% | 1.52% | 0.00% | 5.04% | 68.60% | 5.86% | 10242 |
| | walk | 1.73% | 0.15% | 0.21% | 0.00% | 0.26% | 0.13% | 97.52% | 256464 |
| Total Count | | 42708 | 31152 | 40176 | 2981328 | 39018 | 13026 | 251022 | 3398430 |

Table 9.4 Integrating SVM-based Classification & Stop Detection Results

9.2.3 Network Matching Algorithm Results

The third phase of the framework involves applying network matching as we describe it in chapter 8. The results reveal 99.63% overall accuracy identifying all seven modes. Upon removing the walk and stationary from the validation, an accuracy of 94.28% is revealed for the rest of non-walk/non-stop modes. Table 9.5 presents the confusion index of the results after applying the network matching showing a supreme accuracy of higher than 90% for each mode. It appears that network matching has resolved the confusion identified after applying the classification in subsection 9.2.1. As might be noted from the table below, the confusion between car and train modes has been largely resolved when using the train network to verify stages that were initially classified as train yet belonged to the car mode. The detection accuracy of other modes such as cycling has also been enhanced due to the resolution of its confusion with bus and tube.

| | | Classification | | | | | | | Total Count |
|-------------|------------|----------------|--------|--------|------------|--------|--------|--------|-------------|
| | | bus | car | cycle | stationary | train | tube | walk | |
| Actual | bus | 92.85% | 2.93% | 0.96% | 0.56% | 0.00% | 2.56% | 0.13% | 22500 |
| | car | 2.99% | 91.26% | 0.79% | 2.74% | 1.38% | 0.60% | 0.24% | 62490 |
| | cycle | 1.79% | 4.40% | 93.33% | 0.30% | 0.00% | 0.00% | 0.18% | 43656 |
| | stationary | 0.00% | 0.00% | 0.00% | 99.99% | 0.00% | 0.00% | 0.00% | 2979546 |
| | train | 0.00% | 0.99% | 0.00% | 0.41% | 96.56% | 2.04% | 0.00% | 23556 |
| | tube | 2.39% | 2.27% | 0.00% | 0.00% | 2.10% | 92.77% | 0.47% | 10296 |
| | walk | 0.06% | 0.03% | 0.00% | 0.00% | 0.07% | 0.12% | 99.72% | 256464 |
| Total Count | | 24042 | 60270 | 41454 | 2981328 | 24030 | 11316 | 256068 | 3398508 |

Table 9.5 Integrated Results of Classification, Stop Detection & Network Matching

This section has described the accuracies achieved after applying each phase of the inference framework to the validation dataset collected for this chapter. This section shows the accuracy improvement of mode detection across most modes due to applying network matching, as well as the relatively high accuracy of stop detection. The next section provides a brief discussion of the results achieved in this section by further analysing them and identifying limitations that this study contains, and different problems the framework has encountered along the way.

9.3 Discussion

This section holds a general discussion of the results presented in the previous section. The section mainly stresses on analysing the effect of applying network matching to the results obtained from classification and stops segmentation on each mode individually. We then discuss general limitations we have encountered and accounted for in the process of producing this research. We also analyse the effect of these limitations and propose response strategies that can be used to extend the work of this research while accommodating for these limitations.

9.3.1 Enhancing Accuracy by using Network Matching

One of the most significant steps of the inference framework is the usage of network matching. The accuracy has increased by more than 23% from 71% to 94.28% for non-walk/non-stop modes. In order to understand this difference with respect to each mode however, Table 9.6 is compiled to show the difference of inference accuracy before and after applying network matching. The table expresses the increase in accuracy after applying network matching in blue for the diagonal. The table reflects that the highest improvement occurred in car mode detection as they were well separated from train tracks when matched to the train network. Bus and tube modes have also experienced an improvement in accuracy due matching them to their respective networks.

| | Classification | | | | | | | Total |
|------------|----------------|--------|--------|------------|---------|--------|--------|--------|
| | bus | car | cycle | stationary | train | tube | walk | Count |
| Actual bus | 16.01% | -1.05% | -7.22% | 0.01% | -0.84% | -6.89% | 0.00% | 3750 |
| car | -13.40% | 45.13% | 0.00% | -0.01% | -22.70% | -4.01% | 0.00% | 10415 |
| cycle | -11.48% | 3.39% | 8.29% | 0.00% | 0.00% | -0.21% | 0.00% | 7276 |
| stationary | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 496591 |
| train | 0.00% | 0.00% | 0.00% | 0.00% | 0.46% | -0.33% | -0.13% | 3926 |
| tube | -13.19% | -1.13% | -1.52% | 0.00% | -2.94% | 24.17% | -1.39% | 1716 |
| walk | -1.67% | -0.12% | -0.21% | 0.00% | -0.19% | 0.00% | 2.20% | 42744 |

Table 9.6 Result Differences of before & after applying the Network Matching Process

Conversely, the decrease in the error of the rest of Table 9.6 (other than the diagonal) is expressed in red and in blue if any increase in error is identified. Similar to any of the previous confusion matrices, the lower left triangle expresses Type I errors, which are the incorrect rejection of true null hypothesis (or false positives). On the other hand the upper right triangle expresses Type II errors which are the failure to reject a false null hypothesis (or false negatives). It is clear from Table 9.6 that many car, cycle and tube trips that were initially

classified as bus were cleared into their respective categories after they were rejected from the bus category during the network matching process.

Table 9.7 groups all the reductions in Type I and Type II errors together and adds a third column as the total reduction of both errors. As a result, the table expresses reductions in error after applying network matching in red and any increase in error in blue. It is quite clear that almost the whole table is in red, reflecting the reduction in error for all modes after map matching. It is also apparent that Type I errors for car, train and bus modes were significantly reduced, that is to say; falsely identifying other modes as car, train and bus modes has been reduced. On the other hand, Type II errors for bus and train modes were significantly reduced as a result of network matching, that is to say falsely identifying bus and train modes as other modes has been reduced.

| Mode | Difference in Type I Errors | Difference in Type II Errors | Total Difference |
|----------------|-----------------------------------|------------------------------------|---------------------|
| bus | -16.01% | -43.08% | -60.75% |
| car | -45.13% | 1.09% | -44.04% |
| cycle | -8.29% | -8.96% | -17.25% |
| stationary | 0.00% | -0.01% | -0.01% |
| train | -0.46% | -26.67% | -27.13% |
| tube | -24.17% | -11.44% | -35.62% |
| walk | -2.20% | -5.52% | -7.72% |
| Average | -16.04% | -15.76% | -32.09% |

Blue (+ve) values demonstrate SVM's excellence

Red (-ve) values demonstrate SVM+NM's excellence

Table 9.7 Reduction in Type I/II Errors as a result of using Network Matching

9.3.2 External Limitations

There were several secondary remarks regarding issues briefly mentioned in section 9.1.6 regarding the scope of this research. There are several modes that have not been accounted for in this research due to their irrelevance to the main scope of this thesis. The overground mode for example, has been excluded from the validation results because it was not considered in the inference network due to the absence of its network dataset from the resources available to this research, and hence cannot be verified in the network matching process.

Air travel has also been excluded as there was no training data initially for it in the pilot dataset. However, air travel possesses significantly distinctive speed characteristics, and hence, is relatively easy to detect. The same applies to travelling by ferry where there is a massive lack of ferry data. In fact, the whole dataset contains about 2 return journeys using a ferry.

Another issue with data outside London (or the UK), which is the study area of this research, is that tube and bus network matching cannot be applied due to the lack of network resources. Moreover, speeds of buses in another city (or across cities) will have different movement patterns and with the lack of learning data for these instances these would be relatively difficult to detect.

Table 9.8 below shows how the inference framework before applying network matching has classified GPS data for overground, air and ferry modes and any non-walk mode outside of

London area. The table demonstrates that overground has mostly been classified as car and tube modes due to similarities between their speed values. The ferry trips were all classified as car mode because they usually followed long car journeys and also have similar speeds. Air travel has been mainly spread across car and bus modes. Most travel outside of London has been classified as stationary due to the dominance of stops in any GPS track, while the rest has been classified mainly as tube mode.

| | | Classification | | | | | | Total Count |
|-------------|------------|----------------|---------|------------|--------|-------|-------|-------------|
| | | bus | car | stationary | tube | walks | cycle | train |
| Actual | overground | 0.39% | 73.11% | 3.29% | 17.21% | 0.58% | 2.13% | 3.29% |
| | abroad | 1.18% | 1.67% | 85.97% | 8.85% | 1.23% | 1.10% | 0.00% |
| | air | 45.74% | 53.54% | 0.73% | 0.00% | 0.00% | 0.00% | 0.00% |
| | ferry | 0.00% | 100.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Total Count | | 2544 | 6192 | 74226 | 8166 | 1074 | 1014 | 102 |
| | | | | | | | | 93318 |

Table 9.8 External Transport Modes classified using SVM & Stop Detection

Furthermore, Table 9.9 shows how these modes were categorised after applying network matching to the initial classification of these GPS segments. A major reduction of misclassifying any of these modes into bus and tube modes was corrected due to the matching process. Ferry, overground and air travel has mainly been shifted to be classified as car mode. This makes it easy to integrate these modes in the future, since overground movement can be separated from car by matching to its network, ferry travel by detecting offshore movement, and air travel by means of SVM classification due to its extremely high speed values.

| | | Classification | | | | | | Total Count |
|-------------|------------|----------------|---------|------------|--------|-------|-------|-------------|
| | | bus | car | stationary | tube | walks | cycle | train |
| Actual | overground | 0.00% | 74.27% | 3.29% | 16.44% | 0.58% | 2.13% | 3.29% |
| | Abroad | 0.00% | 11.60% | 85.97% | 0.00% | 1.33% | 1.10% | 0.00% |
| | Air | 0.00% | 98.91% | 0.73% | 0.36% | 0.00% | 0.00% | 0.00% |
| | Ferry | 0.00% | 100.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| Total Count | | 0 | 16290 | 74226 | 522 | 1164 | 1014 | 102 |
| | | | | | | | | 93318 |

Table 9.9 External Transport Modes classified using SVM & Stops Detection followed by Network Matching

Another identified limitation noted from the obtained results is the confusion of cycle and car modes. This confusion stems from classification stage where they both have similar speeds within London. A potential solution to this problem could be separating car travel into two categories, one for travel within congestion zones and another for across towns and cities when using higher level roads such as motorways and A-roads.

9.4 Summary

This chapter demonstrated the performance of the inference framework developed in this thesis to detect the mode of transport from sparse GPS data. The performance is assessed based on a validation dataset that was collected in this chapter for the purpose of assessing the performance of the algorithms developed in this research. The dataset contains data for 84 people over the period from 24/May/2011 to 16/Mar/2012. We have confirmed the suitability of the dataset since it adheres to the specifications we set in chapters 4 and 5 to ensure a

validation process of suitable quality. We have also analysed the dataset in terms of its socio-demographics, device handling and movement characteristics and levels of accessibility to different modes of transport. We have also assessed the effect of these elements on the quality of the validation process. We have also briefly analysed speeds across all participants which seemed to be adequately spread. The geographic extent of the data also highlighted the fact that some modes and activities that occur outside of London will be considered out of the scope of the study mainly due to lack of their relevant network datasets.

We have then presented the validation results for different phases of the inference framework. The moving window SVM-based classification algorithm seemed to achieve similar accuracies when using speed or acceleration as an independent variable in the classification. The results were then aggregated from both variables achieving a relatively satisfactory accuracy yet some confusion between bus and tube and other modes was noted. Almost a quarter of car mode GPS segments were also confused as train mode. Results from the stop detection algorithm were then integrated with the classification results reflecting an accuracy of 99.99% for stop detection. The confusions noted earlier from the classification phase appeared to get resolved by applying the network matching algorithm later in the chapter. After applying network matching, the results reveal **99.63%** overall accuracy for identifying all seven modes including the stationary mode (stops). Upon removing the walk and stationary from the validation, an accuracy of **94.28%** was revealed for the rest of non-walk/non-stop modes.

The final section of the chapter discussed limitations we have encountered while applying the inference algorithm to the GPS validation dataset collected for this research. Among these limitations is the dismissal of other modes such as air, ferry and London Overground travel. This is due to the lack of network information of the London Overground to perform network matching, and the air and ferry modes being out of London's geographical extent; and hence out of the scope of the research. However, when applying the entire inference framework on this subset of the data, we have separated these trips from fast modes such as bus, tube, and train into the car mode category. This was recognised as an advantage where these modes can potentially be easily integrated in future work, since overground movement can be separated from the car category by matching to its network, ferry travel by detecting offshore movement, and air travel by means of SVM classification due to its extremely high speed values. Another identified limitation found from results of this research is the confusion of cycle and car modes as a result of the classification phase where they both have similar speeds inside London. A potential solution to this problem was identified as separating car travel into two categories, one for travel within congestion zones and another when using higher level roads for cross-country travel.

Chapter 10

Conclusions & Further Work

10 CONCLUSIONS & FURTHER WORK

This chapter revisits the research aim and questions previously set in chapter 1 while discussing how this thesis has addressed each of these questions. It also discusses the implications of this research on current policies and standards in place. This is followed by a discussion of the limitations and recommendations for further research work that extends the efforts of this thesis. The chapter finalises this discussion by concluding the achievements and main contributions of work done in this thesis to the research community.

10.1 Introduction

Chapter 1 of this thesis demonstrated the importance of research that aims to infer trip information from movement data to facilitate completing travel surveys that help informing policy makers and transport planners to take better informed decisions. The study was set out to explore the extraction of trip information such as the transport mode from GPS data which stems from the need to obtain less problematic travel information. The main aim of the work was to develop a method that automatically detects the transport mode as a means of understanding human travel behaviour from sparse GPS data. The study has also sought to explore the feasibility of developing the method robust enough to work with no information but raw GPS data in order to reduce participant burden during the process.

The general theoretical literature on this subject is inconclusive on several vital questions within the data and classification method spaces. The study sought to answer four of these questions:

- *What best practices constitute the optimum standards of positional data collection and pre-processing which will ensure obtaining high quality data and a reliable performance validation process for movement-focused travel diaries?*
- *What could constitute the characteristics of an ideal method to detect the transport mode of trip stages (as either stationary, walk, cycle, bus, car, train or underground)? And which independent variable(s) best discriminate between different classes (modes) in this classification problem?*
- *Would breaking each trip into distinct stages, each consisting of a group of consequent GPS segments of the same mode of transport, enhance the accuracy achieved by this classification? Could this also help identify stops and gaps within a GPS track; paving the way to further inferences?*
- *What would be the effect of matching GPS fixes to their respective transport networks on such classification? And what type of information would such method require?*

The following section debates the impact of the work done in this thesis on answering these research questions by discussing the thesis empirical findings and the resulting theoretical implications.

10.2 Findings & Theoretical Implication

The first research question is partially constructive and partially empirical in nature. This question is addressed in chapters 4 and 5 (Data Collection and Data Pre-Processing). The three remaining research questions constitute the methodological questions of this problem. The main methodological respective findings in this research are empirical in nature and chapter-specific; and were summarized within the respective chapters: (Mode Classification, Segmentation and Map Matching). This section will synthesize the empirical findings to answer the study's research objectives.

10.2.1 Research Question 1

1. ***What best practices constitute the optimum standards of positional data collection and pre-processing which will ensure obtaining high quality data and a reliable performance validation process for movement-focused travel diaries?***

The thesis has demonstrated that these best practices could be divided into data collection and pre-processing phases. The case of this research has provided an example of how data collection and pre-processing best practices can be addressed for such travel survey studies. The large amounts of survey types that exist nowadays need to be revisited similarly in order further understand the impact of optimising the data collection standards for these survey types.

Data collection best practices are purpose-specific:

Data collection best practices were conceptualised in this research to include the device type, study area, sample collection rate and sample spatial/temporal and modal distribution. From discussions in the thesis of the first two issues (the choice of the device and the study area), we can conclude that they entirely depend on the purpose of the study. For example studies that are solely interested in places visited by the user for the purpose of calculating the number of household trips might not require long battery live or a highly complex network-dense environment.

The remaining two elements (sample distribution and sample collection rate) also depend on the purpose of the study; however, they will also require some form of analysis or statistical evaluation. For the purpose of this particular study, we have conducted an experiment to identify the **best collection rate** that is suitable to represent movement in London which was found to be around 1 minute. This was considered a research gap that exists in preparation for these types of surveys.

We also highlighted that most studies that aimed to infer the mode of transport from GPS data lacked analysing the adequacy of the **distribution** of the collected samples in the spatial and temporal domains. These studies also lacked accounting for weekly seasonal variation which creates a bias towards that specific week of collection. As a result, we have provided detailed analysis on the collected sample's distribution across space, activities (modes) and time and the duration of GPS data needed to collect representative movements in London. Understanding the data distribution along these domains highlights the limitations that need to be accounted for when reusing and analysing results from this study.

As a result, we collected a GPS dataset for 84 people for duration of 2 months over the period from 24/May/2011 to 16/Mar/2012 at a collection rate of 30 seconds to allow a certain level of redundancy. The geographic extent of the data also highlighted the fact that some modes and activities that occur outside of London will be considered out of the scope of the study mainly due to lack of their relevant network datasets. Ensuring the dataset suitability by adhering to the specifications we set in chapter 4 ensured that the inference framework would have a validation process of suitable quality.

Data pre-processing best practices are case-specific:

We previously highlighted that validation is only as accurate as the metadata provided. This makes pre-processing the collected data an important stage for efficiently assessing the performance of the developed algorithm and for travel surveys in general. Data pre-processing measures do not deviate much from data collection in that they are somewhat depend on the purpose of the study. We have provided a summary of exploratory research findings of data cleansing methods, levels of track generalisation and mode types used in different studies. It was apparent that the most successful studies in urban environments similar to London that applied data cleansing have used speed, acceleration, and altitude filtering methods to address signal loss problems caused by GPS limitations. The context of this study also makes it imperative to include several modes of which London's transport network environment comprises. The level of track generalisation has been also shown to be quite dependant on the modal detail specified for the study. This shows that the best practices set for these pre-processing elements are slightly purpose-dependant but more case-specific to the nature of the study area (being urban or rural) and the complexity of available transport networks.

Another issue handled in this research is finding solutions to the current problems of travel diaries. Many studies similar to research in this thesis do not use participant-reporting to add trip information to GPS tracks and instead depend on assumptions that may bear a large amount of inaccuracy. Inaccurate and under-reporting is also a general problem when it comes to filling GPS-based travel diaries in national surveys as well as in this work as reported from user feedback. The online intelligent travel diary developed in chapter 5 attempted to provide an example of addressing this problem and even provided an incentive by making information available back to the participant. Details of this approach are still specific to the technology-readiness of the audience participating in such studies/surveys. A best practice guide on such approach is an experience that is to yet to be learnt by understanding different cases in which such applications are realised bearing in mind that these approaches are still rather young in this research domain.

10.2.2 Research Question 2

2. *What could constitute the characteristics of an ideal method to detect the transport mode of trip stages (as either stationary, walk, cycle, bus, car, train or underground)? And which independent variable(s) best discriminate between different classes (modes) in this classification problem?*

Ideal method characteristics were found to be mainly pattern-understanding, easily-trained, a priori knowledge-independent and temporally-independent: The characteristics of an ideal method/framework include that it should have a broad understanding of patterns of a trajectory's motion along different trip stages in order to rightfully understand modal behaviours. The method should also be easily trained avoiding a feature extraction process to be able to be potentially extended to function in live dynamic environments. Such method must be capable to work if no information but positional data to be applicable to be used in cases of crowd sourcing online and/or online mass surveys. It should also avoid using any temporal assumptions to ensure the robustness of the method over different samples. As a result, a moving window SVM-based classification algorithm was developed as an example of this idealistic method to infer the mode of transport from sparse GPS data. The classification algorithm revealed 98.24% general accuracy in detecting all modes of transport and 71.05% accuracy for non-walk travel. The confusion resulting from this classification was mainly between bus and tube modes with other modes. Almost a quarter of car GPS segments were also confused as train mode which reflected the need to apply network matching to resolve this confusion.

Most significant independent variables were found to be speed and acceleration: This was shown from outcomes of a statistical evaluation we implemented for the purpose of modal classification which extends on work done in similar studies. Each of the two variables proved to be a better discriminant for certain modes, and therefore, we used both variables in the classification and later applied a selection matrix to produce an integrated classification outcome.

10.2.3 Research Question 3

3. *Would breaking each trip into distinct stages, each consisting of a group of consequent GPS segments of the same mode of transport, enhance the accuracy achieved by this classification? Could this also help identify stops and gaps within a GPS track; paving the way to further inferences?*

Breaking tips into distinct stages does not enhance the accuracy of the classification, however it helps identify exact switch points, stops and gaps in a track: A spatio-temporal clustering method was developed using a combination of travel distance and speed thresholds to identify **stops** within a GPS track while accounting for outliers. Long stays were also detected using a k-means clustering algorithm. Detected stops and long stays were then merged into the classification results in the inference framework. The results revealed accuracy of 99.99% for stops detection slightly enhancing results of some transport modes, yet not affecting much the results of the rest.

10.2.4 Research Question 4

4. *What would be the effect of matching GPS fixes to their respective transport networks on such classification? And what type of information would such method require?*

Network matching significantly helps enhance the accuracy of the classification, and its application will ideally require geometrical and topological network information: Network matching was intended to be applied to transport networks in London which have routing properties (identified routes) such as the underground, bus and train networks. The developed algorithm verified whether non-walk stages follow any of the three network datasets. Part of the algorithm also reasoned about tunnel travel in cases of loss of GPS coverage when travelling on covered areas of the London Underground network. The inference results after applying network matching revealed **99.63%** overall accuracy for identifying all seven modes including the stationary mode (stops) and **94.28%** for the rest of non-walk/non-stop modes. This means that the increase in accuracy as a result of network matching is 23.23%, which reflects the importance of the matching process in resolving any remaining confusions.

10.3 Policy implication

On the governance-level of national travel survey studies, policies have to take into account these set best practices (or standards) to minimise cost and maximise benefit. Moreover, these standards have to be tailored for different survey-types to suit the needs of their respective purposes. Data-related processes such as the device type, sample collection rate, sample distribution, choice of modes and trip elements generalisation level depend very much on the intended purpose of such surveys. Therefore, good governance of such processes and unifying standards by which this travel information is collected and processed would provide a common ground on which periodic data could be compared and analysed for changes and for identifying new trends or responses to new regulations.

The overall advantage of the inference framework developed in this thesis is that it provides a useful tool to make sense out of GPS-based travel surveys by automatically extracting travel information from GPS data. This tool can be used to enrich periodic reports produced by transportation agencies to reveal modal choice/split figures of a higher accuracy. This helps providing a more accurate picture of current transport problems which helps policy makers and transport planners take better informed decisions.

Currently, TfL produces an annual London Travel Demand Survey (LTDS) report which summarises London's travel trends by collecting information from interviews of 8,000 randomly-selected households in London and the surrounding area (TfL, 2013b). The LTDS includes household, individual and trip sheet questionnaires from these households, the latter gathering data on all trips made on a designated travel day (TfL, 2014). As described in chapters 1 and 2, trip sheets tend not to have high response rates and on many occasions may contain inaccurate information. Therefore, the implications of this research (and similar initiatives) on surveys such as LTDS would be enhancing the accuracy of such information and minimising the cost, time and burden of getting this feedback. And in the case of using online intelligent travel diaries such as GeoTravelDiary, feedback could be collected from participants not only with minimal burden but even providing an incentive by sharing back with the participants their own tracks. This study has also provided evidence showing the effectiveness of such tool compared to a paper sheet travel log.

The emergence of new similar applications in the future may give birth to a new way by which more up-to-date travel information is used and shared within a city, country or even across several countries. The future of more frequent travel information updates using these intelligent applications is even more eminent if international organisations/working-groups such as the Open Geospatial Consortium (OGC) took an initiative to standardise the exchange of such big data-generated information via web technologies. Current examples in parallel application spaces already exist; for example sharing up-to-date traffic information via web server end-points such as Google, Waze, and many others. The speed by which information could be contributed online nowadays via smart mobile phones also paves the way for more "open" directions of sharing such information. The wide use of crowdsourcing from travel diary applications such as Moves and Geospike or sport applications such as Garmin Connect and Strava also demonstrates the potential of adopting such technologies to replace traditional travel surveys and make use of the intelligence of these open technologies.

10.4 Limitations and Recommendation for Further Research

The study has offered an evaluative perspective on data properties in the context of its collection and its pre-processing. The study however has encountered a number of limitations, which need to be considered when evaluating the results of the transport mode classification framework.

Dataset limitations include the occurrence of some concentrations of movement in North and West London in the area of Outer London. Outer London also holds a slight spatial representation of bus and cycle modes. Data clusters also seem to frequently occur around transportation hubs. The exhibition of certain modes at specific areas during certain periods of the day only reflects the pattern of life and typical general movement trends that a city would have.

One challenge from the process of data collection was the difficulty of retrieving the GPS devices or ending the study earlier when dealing with younger participant groups. On the other hand, older age groups collected more GPS data in the smaller periods they had the devices for. These observations do not impose a direct threat on the validation performance of the inference framework, yet they are important issues to account for beforehand during the survey planning phase.

Participants' responses from the feedback of the data collection experiment exhibited a strong need for some form of intelligent travel diary application due to the high reported burden levels. An important recommendation is exposing the participants to track maps following the data collection period rather than during it in order to avoid bias. Another recommendation was to use tracking devices with a longer battery life in order to decrease the burden level of device-related tasks.

One of the limitations we have encountered when applying the inference algorithm to the GPS validation dataset collected for this research is the confusion of cycle and car modes as a result of the classification phase where they both have similar speeds inside London. A potential solution to this problem is to separate car travel into two categories, one for travel within London and another when using higher level roads for cross-country travel. This location based-classification could also be extended to other modes such separating bus travel inside London from cross-country coach travel for example.

Another limitation noted from the results of the inference process is the dismissal of other modes such as air, ferry and London Overground travel. The dismissal of the overground travel was due to the lack of network information of the London Overground to perform network matching. Air and ferry modes were dismissed due to their occurrence outside of London and the UK's geographical extent which is the scope of this research. However, when applying the inference framework to GPS data from these modes, these trips were classified as car travel since they did not match to any of the existent transport networks. This was recognised as an advantage where these modes can easily be integrated in future work, since overground movement can be separated from the car category by matching to its network, ferry travel by detecting offshore movement, and air travel by means of SVM classification due to its extremely high speed values.

The inference framework achieved a good level of accuracy across all modes using the guidelines developed in chapter 4. Nevertheless, the accuracy of the device being used and its firmware appeared to have some effect on the classification results which could be explored in further work. Moreover, different rate of GPS data collection could be compared to perform this classification.

Further research can also use outcomes from this research such as stops and indoor detection algorithms to infer significant locations to the participant by reasoning about the stop durations and the repetitiveness of the occurrence of these stops. This is already a well-established line of research as can be found in work done by Bohte & Maat (2009) and Schönfelder & Samaga (2003); however, this thesis provides a solid starting point of drawing these inferences.

Other outcomes from this study can be used in several other applications that can make use of such innovation. For example, the inference can be used to inform Demographic and Health Surveys. Studies such as Dill (2009) that aims to detect whether cycling helps adults meet the recommended levels of physical activity can be directly informed using the type of inference in this research. The usage of results from processed GPS surveys improves data quality, allows meaningful comparisons across studies and populations, and advances health studies field more rapidly (Kerr, et al., 2011).

As mentioned in the previous section, crowdsourcing from different travel diary applications and sport applications paves the way to adopt such smart mobile phone technologies to replace traditional travel surveys. Other research building up on work such as this thesis is widely evolving for example in the University of Leeds where a new direction is being adopted when dealing with Twitter data to use activity space rather than residential space when researching population aspects such as crime, health, etc. (Malleon & Andresen, 2014; Harland, et al., 2014). These emerging research directions in turn will bear many research challenges in the context of management and conflation of the resulting big data. Problems of big data include the capture, organisation, storage, search, sharing, transfer, analysis and visualization of data (Mayer-Schönberger & Cukier, 2014). In the context of the type of data collected in this work, common standards need to be developed to enable common use of such information in a meaningful way to different systems and several research opportunities evolve as a result of this need.

10.5 Conclusion

The work carried out in this thesis has shown that a method that attempts to understand the detailed structure of modal movements could be used to classify sparse GPS into transport modes with no prior knowledge. The moving window SVM-based network-matching classification framework we developed has achieved a superb accuracy of nearly 95% for identifying non-ancillary modes. These results surpass most of achieved accuracies in similar studies for this classification problem-type. As part of the framework, a spatio-temporal clustering method has been developed for identifying stops, gaps within a track achieving an accuracy of 99.99% detection. A network matching method has also been implemented contributing to the framework with an accuracy raise of 25% and paving the way for work aimed at understanding network loads and movements. The framework is also intended for reuse in other cities provided an understanding of the nature of the study area and the complexity of available transport networks.

BIBLIOGRAPHY

- Alpaydin, E., 2004. *Introduction to Machine Learning*. 1st ed. Massachusetts, USA: Massachusetts Institute of Technology Press.
- Anderson, T., Abeywardana, V., Wolf, J. & Lee, M., 2009. *National Travel Survey GPS Feasibility Study: Final Report*, London, UK: Department of Transport.
- ArcGIS, 2012. *Python for ArcGIS*. [Online]
Available at: <http://resources.arcgis.com/en/communities/python/>
[Accessed 30 October 2012].
- Ashbrook, D. & Starner, T., 2003. Using GPS to Learn Significant Locations and Predict Movement Across Multiple Users. *Personal and Ubiquitous Computing*, 7(5), pp. 275-286.
- Bachman, W., Oliveira, M., Xu, J. & Sabina, E., 2012. *Using Household-Level GPS Travel Data to Measure Regional Traffic Congestion*. Washington, D.C., USA, Transportation Research Board 91st Annual Meeting.
- Bloomberg, M. R. & Burden, A. M., 2006. *New York City Pedestrian Level of Service Study: Phase I*, City of New York: NYC Department of City Planning.
- Bohte, W. & Maat, K., 2009. Deriving and Validating Trip Purposes and Travel Modes for Multi-day GPS-Based Travel Surveys: A Large Scale Application in the Netherlands. *Transport Research Part C: Emerging Technologies*, 17(3), pp. 285-297.
- Bolbol, A. & Cheng, T., 2010. *GPS Data Collection Setting For Pedestrian Activity Modelling..* London, UCL.
- Bolbol, A., Cheng, T. & Paracha, A., 2010. *GEOTRAVELDIARY: Towards Online Automatic Travel Behaviour Detection*. Como, Italy, Politecnico di Milano.
- Bolbol, A., Cheng, T. & Tsapakis, I., 2013. *Matching GPS Data to Transport Networks*. University of Liverpool, Liverpool, UK, Geographic Information Science Research in the UK Conference.
- Bolbol, A., Cheng, T., Tsapakis, I. & Chow, A., 2012b. *Sample Size Calculation for Studying Transportation Modes from GPS Data*. Athens, Greece, Procedia - Social and Behavioral Sciences, Elsevier, pp. 3040-3050.
- Bolbol, A., Cheng, T., Tsapakis, I. & Haworth, J., 2012a. Inferring Hybrid Transportation Modes from Sparse GPS Data Using a Moving Window SVM Classification. *Computers, Environment and Urban Systems*, 36(6), p. 526-537.
- Bricka, S. & Bhat, C. R., 2006. Comparative Analysis of Global Positioning System-Based and Travel Survey-Based Data. *Transportation Research Record*, 1972(2006), pp. 9-20.
- Burges, C. J., 1998. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), pp. 121-167.

- Carletta, J., 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, II(22), p. 249–254.
- Chon, H. D., Jun, S., Jung, H. & An, S. W., 2004. Using RFID for accurate positioning. *Journal of Global Positioning Systems*, III(1), p. 32–39.
- Chung, E.-H. & Shalaby, A., 2005. A Trip Reconstruction Tool for GPS-based Personal Travel Surveys. *Transportation Planning and Technology*, 28(5), pp. 381–401.
- Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C., 2009. *Introduction to Algorithms*. 3rd ed. Massachusetts: Massachusetts Institute of Technology.
- Crammer, K. & Singer, Y., 2002. On the Learnability and Design of Output Codes for Multiclass Problems. *Computational Learning Theory*, 47(2-3), pp. 201–233.
- De Groot, A., 1965. *Thought and Choice in Chess*. 1st ed. The Hague: Mouton & Co Publishers.
- De Jong, R. & Mensonides, W., 2003. *Wearable GPS Device as a Data Collection Method for Travel Research*. Washington, DC, USA, Institute of Transport Studies Working Papers.
- Dempster, A. P., Laird, N. M. & Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society*, B(39), pp. 1–39.
- DfT, 2011. *National Travel Survey 2011 GPS Pilot: Summary Analysis*, London, UK: Department for Transport.
- Dice Holdings Inc, 2013. *SourceForge.net: Viking*. [Online] Available at: <http://sourceforge.net/apps/mediawiki/viking/> [Accessed 04 03 2013].
- Dill, J., 2009. Bicycling for Transportation and Health: The Role of Infrastructure. *Journal of Public Health Policy*, Volume 30, pp. 95–110.
- Doherty, S., Papinski, D. & Lee-Gosselin, M., 2006. *An Internet-Based Prompted Recall Diary with GPS Activity-Trip Detection: System Design*. Washington, DC, USA, The 85th Annual Meeting of the Transportation Research Board.
- Draijer, G., Kalfs, N. & Perdok, J., 2000. Global Positioning System as Data Collection Method for Travel Research. *Transportation Research Record: Journal of the Transportation Research Board*, 1719(1), pp. 147–153.
- Ebbinghaus, H., 1885/1913. *Memory: A Contribution to Experimental Psychology*. 1st ed. New York: Teachers College Press, Columbia University.
- Edwards, D. et al., 2009. *Visitors to urban destinations: understanding tourist 'experiences' and 'behaviour' in cities, an Australian case study*, Australia: Gold Coast, CRC for Sustainable Tourism Pty Ltd..
- Everitt, B. S. & Dunn, G., 2010. *Applied Multivariate Data Analysis*. 2nd ed. London, England: Edward Arnold.

- Forrest, T. L. & Pearson, D. F., 2005. Comparison of Trip Determination Methods in Household Travel Surveys Enhanced by a Global Positioning System. *Transportation Research Record: Journal of the Transportation Research Board*, Volume 1917, pp. 63-71.
- Gallagher, T., Li, B., Kealy, A. & Dempster, A., 2009. *Trials of Commercial Wi-Fi Positioning Systems for Indoor and Urban Canyons*. Gold Coast, Australia, International Global Navigation Satellite Systems Society IGNSS Symposium.
- Garmin Inc., 2011. *GPS 35 LP TracPak GPS Smart Antenna Technical Specification*. [Online] Available at: <http://www.garmin.com/products/gps35/> [Accessed 17 February 2011].
- GlobalMotion Media Inc, 2013. *EveryTrail - Travel Community, iPhone Guides for Sightseeing, Hiking, Walking Tours and More*. [Online] Available at: <http://www.everytrail.com/> [Accessed 04 03 2013].
- Gong, H., Chen, C., Bialostozk, E. & Lawson, C. T., 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36(2), pp. 131-139.
- Goodchild, M. F., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, Volume II, p. 24-32.
- Gray, P., 2001. *Psychology*. 4th ed. New York, USA: Worth Publishers.
- Griffiths, R., Richardson, A. J. & Lee-Gosselin, M. E. H., 2000. *Travel Surveys, Transportation in the New Millennium*. Washington, DC, USA, Transportation Research Board 79th Annual Meeting.
- GTrek, 2012. *GTrek GPS Data Logger and Tracker*. [Online] Available at: <http://www.gtrek.co.uk> [Accessed 27 10 2012].
- Guensler, R. L. et al., 2006. *Analysis of Commute Atlanta Instrumented Vehicle GPS Data Destination Choice Behavior and Activity Spaces*. Washington, D.C., USA, Transportation Research Board 85th Annual Meeting.
- Hacklay, m. & Weber, P., 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, December, 7(4), pp. 12-18.
- Harland, K., Birkin, M. & Martin, D., 2014. *Simulating the Pulse of the City: Using a Combined Approach for Modelling Commuter Patterns*. Glasgow, GIS Research in UK, University of Glasgow.
- Hinch, S. W., 2007. *Outdoor Navigation with GPS*. 2nd ed. Berkeley: Wilderness Press.
- Horthersall, D., 1995. *History of Psychology*. 3rd ed. New York: McGraw-Hill.

Howe, J., 2006. *Wired 14.06: The Rise of Crowdsourcing*. [Online]
Available at: <http://www.wired.com/wired/archive/14.06/crowds.html>
[Accessed 26 April 2010].

Innersource, 2013. *GPS Tracks and Maps*. [Online]
Available at: <http://maps.innersource.com/>
[Accessed 04 02 2013].

Jiang, B., Yin, J. & Zhao, S., 2009. Characterizing the Human Mobility Pattern in a Large Street Network. *Physical Review E: statistical, nonlinear, and soft matter physics*, 80(2), p. 021136.

Jun, J., Guensler, R. L. & Ogle, J. H., 2006. Smoothing Methods to Minimize Impact of Global Positioning System Random Error on Travel Distance, Speed, and Acceleration Profile Estimates. *Transportation Research Record: Journal of the Transportation Research Board*, 1972(1), pp. 141-150.

Kaufman, L. & Rousseeuw, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. 1st ed. s.l.:John Wiley & Sons.

Kerr, J., Duncan, S. & Schipperjin, J., 2011. Using Global Positioning Systems in Health Research: A Practical Approach to Data Collection and Processing. *American Journal of Preventive Medicine*, 41(5), pp. 532-540.

Klecka, W. R., 1980. *Discriminant Analysis*. Beverly Hills, California, US: Sage Publications, Inc.

Knoblauch, R. L., Pietrucha, M. T. & Nitzburg, M., 1996. Field Studies of Pedestrian Walking Speed and Start-Up Time. *Transportation Research Record*, Issue 1538, pp. 27-38.

Landis, J. R. & Koch, G. G., 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, March, 33(1), pp. 159-174.

Liao, L., Fox, D. & Kautz, H., 2007. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *The International Journal of Robotics Research*, 26(1), pp. 119-134.

Li, Z. J. & Shalaby, A. S., 2008. *Web-Based GIS System for Prompted Recall of GPS-Assisted Personal Travel Surveys: System Development and Experimental Study*. Washington, DC, USA, Transportation Research Board 87th Annual Meeting.

MacQueen, J., 1967. Some Methods for Classification and Analysis of Multivariate Observations. *The 5th Berkeley Symposium of Mathematical Statistics*, Volume 1, pp. 281-297.

Malleson, N. & Andresen, M., 2014. *Using Social Media Data to Assess Spatial Crime Hotspots*. Glasgow, UK, GIS Research in UK, University of Glasgow.

Manzoni, V., Maniloff, D., Kloeckl, K. & Ratti, C., 2010. *Transportation Mode Identification and Real-Time CO₂ Emission Estimation Using Smartphones*, Cambridge, MA, US: SENSEable City Lab.

- MapmyFitness Inc, 2013. *Map Fitness Training and Track Fitness Workouts* | MapMyFitness. [Online]
Available at: <http://www.mapmyfitness.com>
[Accessed 04 03 2013].
- Mayer-Schönberger, V. & Cukier, K., 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. 1st ed. Boston, USA: Eamon Dolan/Houghton Mifflin Harcourt.
- Mitchell, T., 1997. *Machine Learning*. 1st ed. s.l.:McGraw Hill.
- Montello, D. R., 2001. Spatial Cognition. In: N. J. Smelser & P. B. Baltes, eds. *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Pergamon Press, pp. 14771-14775.
- NTS, 2011. *Department for Transport: A Public Consultation on Future Design of the National Travel Survey*, London: Department for Transport.
- Ortúzar, J. d. D. & Willumsen, L. G., 2011. *Modelling Transport*. 4th ed. London, England: John Wiley & Sons Inc..
- OSM, 2012. *OpenStreetMap*. [Online]
Available at: <http://www.openstreetmap.org/>
[Accessed 30 October 2012].
- Perotto-Baldivieso, H., Cooper, S. M., Figueroa-Pagan, M. & Romo, J., 2008. *Too Much Data? Optimizing GPS Collar Data Collection Schedules*. Louisville, KY, USA, The 2008 Joint Meeting of the Society for Range Management and the America Forage and Grassland Council.
- Pyle, D., 1999. *Data Preparation for Data Mining*. 1st ed. Los Altos, CA, USA: Morgan Kaufmann Publishers.
- Quddus, M., 2006. *High Integrity Map Matching Algorithms for Advanced Transport Telematics Applications*, London: Imperial College London.
- R Project, 2012. *The R Project for Statistical Computing*. [Online]
Available at: <http://www.r-project.org/>
[Accessed 29 October 2012].
- R.A. Malatest & Associates Ltd., 2012. *2011 CRD Origin-Destination Household Travel Survey: Daily Travel Characteristics Report*, Canada: Malatest.
- Reddy, S. et al., 2010. Using Mobile Phones to Determine Transportation Modes. *ACM Transactions on Sensor Networks (TOSN)*, VI(2), pp. 1-27.
- Ren, M. & Karimi, H. A., 2009. A Chain-Code-Based Map Matching Algorithm for Wheelchair Navigation. *Transactions in GIS*, 13(2), p. 197-214.
- Rodrigue, J. P., Comtois, C. & Slack, B., 2006. *The Geography of Transport Systems*. 1st ed. New York, USA: Routledge.

Roorda, M. J. & Miller, E. J., 2004. *Toronto Activity Panel Survey - A Multi-Instrument Panel Survey*. Playa Herradura, Costa Rica, ISCTSC.

Schneider, A., 2013. *GPS Visualizer*. [Online]
Available at: <http://www.gpsvisualizer.com/>
[Accessed 04 03 2013].

Schöenfelder, S., Axhausen, K., Antille, N. & Bierlaire, M., 2002. Exploring the Potentials of Automatically Collected GPS Data for Travel Behaviour Analysis - A Swedish Data Source. *GI-Technologien für Verkehr und Logistik*, Volume 13, pp. 155-179.

Schöenfelder, S. et al., 2006. *Analysis of Commute Atlanta Instrumented Vehicle GPS Data: Destination Choice Behavior and Activity Spaces*. Washington, DC, USA, Transportation Research Board 85th Annual Meeting.

Schöenfelder, S. & Samaga, U., 2003. *Where Do You Want to Go Today? — More Observations on Daily Mobility*. Ascona, Switzerland, The 3rd Swiss Transport Research Conference.

Schüssler, N. & Axhausen, K. W., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. *Transport Research Record: Journal of Transport Research Board*, Issue 2105, pp. 28-36.

Schwieger, V., 2007. *Positioning within the GSM Network*. San Jose, Costa Rica, FIG Regional Conference, pp. 1-14.

Shawe-Taylor, J. & Cristianini, N., 2000. *Support Vector Machines and Other Kernel-Based Learning Methods*. 1st ed. Cambridge, UK: Cambridge University Press.

Sheskin, D. J., 2004. *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd ed. London: Chapman and Hall/CRC Press.

Simpson, D., 2011. *Route Choice Variability: A reactive or proactive decision?*, London, UK: UCL: Centre of Transport Studies.

Steinwart, I. & Christmann, A., 2008. *Support Vector Machines*. 1st ed. New York, USA: Springer-Verlag.

Stenneth, L., Wolfson, O., Yu, P. S. & Xu, B., 2011. *Transportation Mode Detection using Mobile Phones and GIS Information*. New York, NY, US, ACM Press, pp. 54-63.

Stopher, P., Clifford, E., Zhang, J. & FitzGerald, C., 2008a. *Deducing Mode and Purpose from GPS Data*. Washington, DC, USA, Transportation Research Board.

Stopher, P. R., 2008. *Collecting and Processing Data from Mobile Technologies*. Annecy, France, ISCTSC.

Stopher, P. R. & Collins, A., 2005. *Conducting a GPS prompted recall survey over the internet*. Washington, DC, USA, The 84th Annual Meeting of the Transportation Research Board.

Stopher, P. R. & Greaves, S. P., 2007. Household travel surveys: Where are we going?. *Transportation Research Part A: Policy and Practice*, 41(5), p. 367–381.

Stopher, P. R., Jiang, Q. & FitzGerald, C., 2005. *Processing GPS Data from Travel Surveys*. Toronto, Canada, The Second International Colloquium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications.

Stopher, P. R. & Metcalf, H. M. A., 1996. *Methods for Household Travel Surveys*.. Washington, D.C., USA, National Academy Press.

TfL, 2008. *Distance between adjacent Underground stations - a Freedom of Information request to Transport for London - WhatDoTheyKnow*. [Online]

Available at:

https://www.whatdotheyknow.com/request/distance_between_adjacent_underg#incoming-5516

[Accessed 12 July 2014].

TfL, 2009a. *Economic impact of traffic signals*, Greater London: Transport for London in collaboration with London Development Agency.

TfL, 2009b. *Travel in London: Key Trends and Developments Report Number 1*, London: Transport for London.

TfL, 2012a. *Key Facts*. [Online]

Available at:

<http://www.tfl.gov.uk/corporate/modesoftransport/londonunderground/1608.aspx>

[Accessed 23 October 2012].

TfL, 2012b. *London Buses*. [Online]

Available at: <http://www.tfl.gov.uk/corporate/modesoftransport/1548.aspx>

[Accessed 7 November 2012].

TfL, 2012c. *TfL Journey Planner*. [Online]

Available at:

http://journeyplanner.tfl.gov.uk/user/XSLT_SEL_STT_REQUEST?sessionID=o&language=en&mode=line&linePreSel=tf1:25:*&linePreSel=tf1:63:

[Accessed 28 October 2012].

TfL, 2012d. *Tube stations above and below ground - a Freedom of Information request to Transport for London - WhatDoTheyKnow*. [Online]

Available at:

https://www.whatdotheyknow.com/request/tube_stations_above_and_below_gr#incoming-257818

[Accessed 14 July 2014].

TfL, 2013a. *Travel in London: Report 6*, London: Transport for London (TfL).

TfL, 2013b. *Travel in London, Supplementary Report: London Travel Demand Survey (LTDS)*, London: Transport for London Official Report.

- TfL, 2014. *London Travel Demand Survey - Transport for London*. [Online]
Available at: <https://www.tfl.gov.uk/corporate/about-tfl/how-we-work/planning-for-the-future/consultations-and-surveys/london-travel-demand-survey>
[Accessed 01 September 2014].
- Tierney, k. et al., 1996. *Travel Survey Manual Appendices*, Cambridge, MA: Cambridge Systematics, Inc..
- Tolman, E. C., 1948. Cognitive Maps in Rats and Men. *The Psychology Review*, 189(55), pp. 189-208.
- Tsapakis, I. et al., 2012. Effects of Tube Strikes on Journey Times in the Transport Network of London. *Transportation Research Record*, 1(2012), p. 84-92.
- Tsui, S. Y. A. & Shalaby, A. S., 2006. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*, Volume 1972, pp. 38-45.
- u-blox, 2009. *YUMA: Software and Service for Capture and Process*. [Online]
Available at: <http://www.u-blox.com/en/gps-solutions/yuma.html>
[Accessed 29 August 2009].
- UK Census, 2011. *Census Dissemination Unit: Postcode Data*. [Online]
Available at: <http://cdu.mimas.ac.uk/pclut/>
[Accessed 23 11 2013].
- Van den Bossche, F., Wets, G. & Brijs, T., 2005. *The Use of Travel Survey Data in Road Safety Analysis*. Strasbourg, France, The European Transport Conference.
- Vapnik, V. N., 1998. *Statistical Learning Theory*. 1st ed. New York, USA: Wiley-Interscience.
- Wang, H., Calabrese, F., Lorenzo, G. D. & Ratti, C., 2010. *Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records*. Washington, DC, US, IEEE Computer Society.
- Weidmann, U., 1993. *Transporttechnik der Fussg: - Transporttechnische Eigenschaften des Fussgängerverkehrs (Literaturstudie)*, Zürich: Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau IVT an der ETH Zürich.
- Wolf, J., Oliveira, M. & Thompson, M., 2003. Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. *Transportation Research Record: Journal of the Transportation Research Board*, 1854 / 2003(1), pp. 189-198.
- Wood, C. & Mace, O., 2001. Dead Reckoning Keeps GPS in Line. *GPS World*, 12(3), pp. 14-23.
- Xu, Y., 2010. *Effective GPS-Based Panel Survey Sample Size for Urban Travel Behaviour Studies.*, Atlanta, US: Georgia Institute of Technology: School of Civil and Environmental Engineering..

- Yang, D., Cai, B. & Yuan, Y., 2003. *An Improved Map-Matching Algorithm Used in Vehicle Navigation System*. s.l., IEEE Proceedings on Intelligent Transportation Systems, pp. 1246-1250.
- Yang, J.-S., Kang, S.-P. & Chon, K.-S., 2005. The Map Matching Algorithm of GPS Data with Relatively Long Polling Time Intervals. *Journal of the Eastern Asia Society for Transportation Studies*, Volume 6, pp. 2561-2573.
- Yin, H. & Wolfson, O., 2004. *A Weight-Based Map Matching Method in Moving Objects Databases*. Washington, DC, USA, IEEE Computer Society, pp. 437 - 438.
- Zheng, Y. et al., 2010. Understanding Transportation Modes Based on GPS Data for Web Applications. *ACM Transactions on the Web (TWEB)*, IV(1), pp. 1-36.
- Zheng, Y., Liu, L., Wang, L. & Xie, X., 2008. *Learning transportation mode from raw GPS data for geographic applications on the Web*. Beijing, China, ACM Press, pp. 247-256.

APPENDICES

Appendix A: Placement Interview

| | | | | | | | |
|----------------------------------|---|-----------------------|--|---------------------|--|--------------------|--|
| GPS Placement Prompts 1/6 | | | | | | | |
| Forename(s) | | | | | | | |
| Surname | | | | | | | |
| | PROMPT: Has the participant read the GPS Survey Information Sheet | | | | | | |
| | PROMPT: Can the participant explain the study in their own words? | | | | | | |
| | PROMPT: Has the participant signed the Informed Consent Form? | | | | | | |
| | PROMPT: Explain features of GPS device (On, Off, Battery Charger) | | | | | | |
| | PROMPT: Explain how to omit trips - switch off GPS device | | | | | | |
| Age | | 18-24 | | 25-39 | | 40-59 | |
| | | 60+ | | | | | |
| Occupation | | Full Time Employed | | Part Time Employed | | Self Employed | |
| | | Retired | | Paid Work | | Not Working | |
| | | Full-Time Education | | Part-Time Education | | Retired | |
| | | Other (please state): | | | | | |
| Job Title | | | | | | | |
| Hours worked per week | | Less than 4 hours | | 4-8 hours | | 8-18 hours | |
| | | 18-36 hours | | 36-45 hours | | More than 45 hours | |
| Hours worked per week | | Regular Working Hours | | | | | |
| | | Irregular (Shifts?) | | | | | |
| Participant Number | | | | GPS Device Number | | | |
| GPS Placement Date | | | | GPS End Date | | | |
| Data Col. Date & Time | | | | | | | |
| GPS Return Date & Time | | | | | | | |
| GPS Return Location | | | | | | | |
| | PROMPT Voucher selection to receive when GPS device is returned | | | | | | |
| | £10 Voucher | HMV | | iTunes | | Marks & Spencer | |
| | | | | | | | |
| Email | | | | | | | |
| Phone | | | | | | | |

| GPS Placement Prompts 2/6 | | | | |
|-------------------------------|---|--------------------|------|----|
| Modes of Travel | Do you have a driving licence? | | YES | NO |
| | Do you have access to a car? | | YES | NO |
| | Do you have access to a bicycle? | | YES | NO |
| | Are you a member of the Barclays Cycle Hire scheme? | | YES | NO |
| | Did you travel outside of London in the past month? | | YES | NO |
| | Did you travel outside of London in the past 2 months? | | YES | NO |
| | Have you flown by airplane in the past month? | | YES | NO |
| | Have you flown by airplane in the past 2 months? | | YES | NO |
| | Will you travel outside of London in the next 2 months? | | YES | NO |
| | Details: | | | |
| | Will you fly in the next 2 months? | | YES | NO |
| | Details: | | | |
| | Will you make any trips abroad in the next 2 months? | | YES | NO |
| | Details: | | | |
| Thinking about your travel... | In the past 2 months have you travelled by: | | | |
| | Foot/Bicycle | London Underground | Rail | |
| | Car (driver) | Car (passenger) | Van | |
| | Bus | Other: | | |
| | In the past month have you travelled by: | | | |
| | Foot/Bicycle | London Underground | Rail | |
| | Car (driver) | Car (passenger) | Van | |
| | Bus | Other: | | |
| | In the past week have you travelled by: | | | |
| | Foot/Bicycle | London Underground | Rail | |
| | Car (driver) | Car (passenger) | Van | |
| | Bus | Other: | | |
| | Notes | | | |

| GPS Placement Prompts 3/6 | | |
|----------------------------|--|--|
| Household Structure | How many adults live in your home? | |
| | How many people aged under 18 live in your home? | |
| | How many people aged under 16 live in your home? | |
| | PROMPT Identification of destinations and locations travelled for analysis | |
| | PROMPT Home/Friends/Family? | |
| | PROMPT Work/Education? | |
| | PROMPT Other Commitments/Charity Work? | |
| | PROMPT Hobbies/Activities/Sports/Leisure/Gym? | |
| | PROMPT Shopping? | |
| | PROMPT: Leave Card & THANKYOU!! | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| Notes | | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| Notes | | |
| <i>Notes</i> | | |
| | | |
| | | |
| | | |
| | | |

| GPS Placement Prompts 4/6 | | |
|---------------------------|-------------|--|
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| <i>Notes</i> | | |
| | | |
| | | |

| GPS Placement Prompts 5/6 | | |
|---------------------------|-------------|--|
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| Notes | | |
| | | |
| | | |

| GPS Placement Prompts 6/6 | | |
|---------------------------|-------------|--|
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| | Description | |
| | Activity | |
| | Address | |
| | | |
| | | |
| | Notes | |
| Notes | | |
| | | |
| | | |

Appendix B: Exit Interview

| Exit Interview 1/2 | | | | | | | | | | | | | |
|------------------------------------|--|---|---|------------|---|----|---|---|-----------------|----|-----------|----------------|--|
| Carrying the Device | How easy did you find it to keep the device charged? | | | | | | | | | | | | |
| | Very Easy | | | Quite Easy | | OK | | | Quite Difficult | | | Very Difficult | |
| | Did the device ever run out of charge? | | | | | | | | YES | | | NO | |
| | Details: | | | | | | | | | | | | |
| | How easy did you find it to remember to carry the device? | | | | | | | | | | | | |
| | Very Easy | | | Quite Easy | | OK | | | Quite Difficult | | | Very Difficult | |
| | Did you ever forget to carry the device? | | | | | | | | YES | | | NO | |
| | Details: | | | | | | | | | | | | |
| | | | | | | | | | | | | | |
| | Did you tend to forget to carry the device for specific types of trips? | | | | | | | | | | | | |
| | Details: | | | | | | | | | | | | |
| | On a scale of 1-10, how do you rate the burden of carrying the device daily? | | | | | | | | | | | | |
| | LOW | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | HIGH | |
| | Details: | | | | | | | | | | | | |
| | On a scale of 1-10, how do you rate the burden of carrying the device for the eight week survey period? | | | | | | | | | | | | |
| | LOW | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | HIGH | |
| | | | | | | | | | | | | | |
| | On a scale of 1-10, how do you rate the burden of filling the travel diary log daily? | | | | | | | | | | | | |
| LOW | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | HIGH | | |
| Details: | | | | | | | | | | | | | |
| Meeting with the researcher | How easy was it to arrange your initial GPS Placement interview? | | | | | | | | | | | | |
| | EASY | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | DIFFICULT | |
| | Did you find the halfway meeting helpful or a burden? | | | | | | | | | | | | |
| | HELPFUL | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | BURDENSOME | |
| | How easy was it to arrange a meeting to drop off the GPS device? | | | | | | | | | | | | |
| | EASY | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | DIFFICULT | |
| | How easy was it to communicate with the researcher during the survey? | | | | | | | | | | | | |
| EASY | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | DIFFICULT | | |

| Exit Interview 2/2 | | | | |
|-------------------------------|---|--------------|--------------------|----|
| Thinking about your travel... | Over the survey period, have you travelled by: | | | |
| | Foot | Bicycle | London Underground | |
| | Rail | Car (driver) | Car (passenger) | |
| | Bus | Van | Plane | |
| | Other: | | | |
| | Have you made any trips out of the ordinary? | | YES | NO |
| | Details: | | | |
| | Have you moved house? | | YES | NO |
| | Details: | | | |
| | Have you bought a bike/car? | | YES | NO |
| | Details: | | | |
| | Have you changed jobs? | | YES | NO |
| | Details: | | | |
| | Have any other circumstances influenced the way you have travelled? | | | |
| | Details: | | | |
| The project | Would you have been willing to continue for a longer survey period without a reward (i.e. without receiving another gift voucher)? | | YES | NO |
| | Would you have been willing to continue for a longer survey period if you were given another reward (i.e. a further gift voucher) | | YES | NO |
| | If you answered yes above, how much additional reward would you have required to accept to continue the survey for another eight weeks? | | | |
| | Details: | £ | | |
| Notes | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

Appendix C: Paper Travel Diary Log Template

Name

Start Date:

[illegible]