Developing Travel Behaviour Models Using Mobile Phone Data

Andrew Bwambale

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds

Institute for Transport Studies

December 2018

Intellectual property and publications

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in chapters 2, 3, 4, and 5 consists of original research, where the candidate was the lead researcher. The main ideas in these chapters were developed by the candidate during the PhD with the support and guidance of the supervisors and other researchers, who have been listed as co-authors where appropriate.

All the manuscripts were written by the candidate and improved by comments from all the co-authors. The manuscripts are summarised below;

Chapter 2 has been published as follows;

> **Bwambale, A.**, Choudhury, C.F. and Hess, S., 2017. Modelling trip generation using mobile phone data: a latent demographics approach. Journal of Transport Geography (in press)

Chapter 3 was submitted for review to Transportmetrica A: Transportation Science and the candidate has received an invitation to revise the manuscript;

> **Bwambale, A.**, Choudhury, C.F., Hess, S. (under review) Modelling long-distance route choice using mobile phone call detail record data: a case study of Senegal.

Chapter 4 was submitted for review to Transportation Part A: Policy and Practice;

> **Bwambale, A.**, Choudhury, C.F., Hess, S. (under review) Modelling departure time choice using mobile phone data.

Chapter 5 has been prepared for submission to the Journal *Transportation*;

> **Bwambale, A.**, Choudhury, C.F., Hess, S., Iqbal, M. S. (to be submitted) Getting the best of both worlds - a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling.

Acknowledgements

I would like to express my sincere gratitude to my supervisors Charisma F. Choudhury and Stephane Hess for their support, guidance, insightful thoughts and comments on my work. They have inspired me to always aim for the best and created opportunities for me to meet and network with like-minded researchers.

Special thanks go to my examiners Mohammed Quddus and Simon Shepherd who made my viva an interesting discussion from which I learnt new ideas and perspectives about my work and transport research in general. Simon had earlier on assessed my work at the transfer stage and I thank him for the encouraging comments.

I would also like to thank other members of staff and PhD students, particularly those in the Choice Modelling group, who always provided constructive feedback on the work I presented during the various seminars.

This research would not have been possible without the funding of the Economic and Social Research Council (ESRC) of the UK and the Institute for Transport Studies, University of Leeds through the Advanced Quantitative Methods (AQM) scholarship. I am extremely grateful to them for this great opportunity. I also acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS.

The data used in this research was obtained from various institutions and service providers including, the Idiap Research Institute in Switzerland, Sonatel group in Senegal, Orange group in France, and Grameenphone Ltd in Bangladesh. I thank them for making their data available for research purposes.

Finally, I would like to thank my family: my wife Agatha, my children Angelina and Frida, my parents Mr Leo and Mrs Petronilla Bwambale, and my brothers and sisters for standing with me through the last three years, and offering their prayers and words of encouragement.

Abstract

Improving the performance and efficiency of transport systems requires sound decision-making supported by data and models. However, conducting travel surveys to facilitate travel behaviour model estimation is an expensive venture. Hence, such surveys are typically infrequent in nature, and cover limited sample sizes. Furthermore, the quality of such data is often affected by reporting errors and changes in the respondents' behaviour due to awareness of being observed. On the other hand, large and diverse quantities of time-stamped location data are nowadays passively generated as a by-product of technological growth. These passive data sources include Global Positioning System (GPS) traces, mobile phone network records, smart card data and social media data, to name but a few. Among these, mobile phone network records (i.e. call detail records (CDRs) and Global Systems for Mobile Communication (GSM) data) offer the biggest promise due to the increasing mobile phone penetration rates in both the developed and the developing worlds. Previous studies using mobile phone data have primarily focused on extracting travel patterns and trends rather than establishing mathematical relationships between the observed behaviour and the causal factors to predict the travel behaviour in alternative policy scenarios.

This research aims to extend the application of mobile phone data to travel behaviour modelling and policy analysis by augmenting the data with information derived from other sources. This comes along with significant challenges stemming from the anonymous and noisy nature of the data. Consequently, novel data fusion and modelling frameworks have been developed and tested for different modelling scenarios to demonstrate the potential of this emerging low-cost data source.

In the context of trip generation, a hybrid modelling framework has been developed to account for the anonymous nature of CDR data. This involves fusing the CDR and demographic data of a sub-sample of the users to estimate a demographic prediction sub-model based on phone usage variables extracted from the data. The demographic group membership probabilities from this model are then used as class weights in a latent class model for trip generation based on trip rates extracted from the GSM data of the same users. Once estimated, the hybrid model can be applied to probabilistically infer the socio-demographics, and subsequently, the trip generation of a large proportion of the population where only large-scale anonymous CDR data is available as an input. The estimation and validation results using data from Switzerland show that the hybrid model competes well against a typical trip generation model estimated using data with known socio-demographics of the users. The hybrid framework can be applied to other travel behaviour modelling contexts using CDR data (in mode or route choice for instance).

The potential of CDR data to capture rational route choice behaviour for long-distance inter-regional O-D pairs (joined by highly overlapping routes) is demonstrated through data fusion with information on the attributes of the alternatives extracted from multiple external sources. The effect of location discontinuities in CDR data (due to its event-driven nature), and how this impacts the ability to observe the users' trajectories in a highly overlapping network is discussed prompting the development of a route identification algorithm that distinguishes between unique and broad sub-group route choices. The broad choice framework, which was developed in the context of vehicle type choice is then adapted to leverage this limitation where unique route choices cannot be observed for some users, and only the broad sub-groups of the possible overlapping routes are identifiable. The estimation

and validation results using data from Senegal show that CDR data can capture rational route choice behaviour, as well as reasonable value of travel time estimates.

Still relying on data fusion, a novel method based on the mixed logit framework is developed to enable the analysis of departure time choice behaviour using passively collected data (GSM and GPS data) where the challenge is to deal with the lack of information on the desired times of travel. The proposed method relies on data fusion with travel time information extracted from Google Maps in the context of Switzerland. It is unique in the sense that it allows the modeller to understand the sensitivity attached to schedule delay, thus enabling its valuation, despite the passive nature of the data. The model results are in line with the expected travel behaviour, and the schedule delay valuation estimates are reasonable for the study area.

Finally, a joint trip generation modelling framework fusing CDR, household travel survey, and census data is developed. The framework adjusts the scaling factors of a traditional trip generation model (based on household travel survey data only) to optimise model performance at both the disaggregate and aggregate levels. The framework is calibrated using data from Bangladesh and the adjusted models are found to have better spatial and temporal transferability.

Thus, besides demonstrating the potential of mobile phone data, the thesis makes significant methodological and applied contributions. The use of different datasets provides rich insights that can inform policy measures related to the adoption of big data for transport studies. The research findings are particularly timely for transport agencies and practitioners working in contexts with severe data limitations (especially in developing countries), as well as academics generally interested in exploring the potential of emerging big data sources, both in transport and beyond.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Context

Over the last few decades, technological advances have facilitated the storage, retrieval, and processing of extremely large amounts of timestamped location data. These datasets are increasingly becoming important sources of information for transportation studies (Clarke, 2016). The interest in big data is largely driven by its ability to passively capture the behavioural patterns of wider proportions of the population at a low cost compared to traditional data collection methods (such as household surveys), which are usually expensive. The high costs of traditional surveys often lead to small sample sizes and an increased risk of sampling bias. This problem is usually exacerbated by the low response rates due to survey response burden (e.g. Rolstad et al., 2011, Groves, 2006), as well as potential changes in the respondents' behaviour due to awareness of being observed.

Although the challenges with traditional data collection methods can be found in both the developed and the developing worlds, they are more prevalent in the latter context due to stringent budget constraints and limited initiatives to systematically document the few available travel survey records from related studies (San Santoso and Tsunokawa, 2005). Therefore, the emergence of low-cost big data presents a timely opportunity to address some of the data limitations especially in those contexts.

Large-scale travel patterns have previously been extracted from Global Positioning System (GPS) data (e.g. Li et al., 2018, Hess et al., 2015, Broach et al., 2012, Bierlaire et al., 2010, Papinski et al., 2009), mobile network records (e.g. Çolak et al., 2015, Iqbal et al., 2014, Jiang et al., 2013, Schlaich et al., 2010), wireless sensor data (e.g. Tubaishat et al., 2009), bluetooth sensor data (e.g. Crawford, 2017, Hainen et al., 2011), smart card transaction records (e.g. Ma et al., 2013, Munizaga and Palma, 2012, Pelletier et al., 2011, Liu et al., 2009), and social media data (e.g. Hawelka et al., 2014, Hasan et al., 2013) among others. Apart from smart card and social media data, most of the other data types are occasionally referred to as mobile phone data.

However, each data type has limitations. For example, the generation of mobile phone GPS data requires smartphones and regular access to the internet. Wireless and bluetooth data require the installation of sensor infrastructure across the study area to detect the mobile phone movements. While such datasets have been successfully used for transport studies in developed countries, they cover much smaller proportions in the developing world, where mobile internet penetration (e.g. 21% of the population in sub-Saharan Africa) and smartphone adoption (e.g. 34% of all mobile connections in sub-Saharan Africa) is still very low (GSMA, 2018), which would further increase the risk of sampling bias.

To avoid such issues, this research proposes the use of mobile phone network records, which are independent of the phone type and internet accessibility, thereby being more widely available. These records may either be event-driven, such as Call Detail Records (CDRs) and cell phone handover data, which report the timestamped locations of communication events only, or network-driven, such as location-area update, signalling or Global System for Mobile communications (GSM) data, which report the IDs of all the cells traversed by an active mobile phone at regular time intervals (irrespective of the calling or texting

patterns of the users). Tables 1-1 and 1-2 show typical excerpts of CDR and GSM data, respectively.

Table 1-1 Excerpt of the CDR data

Anonymous User ID	Date	Time	Duration	Tower Longitude	Tower latitude
ABH03JACKAAAgfBALA	20120624	11:41:49	25	23.9339	90.2931
ABH03JAC8AAAbZfAHW	20120624	13:43:25	13	23.7931	90.2603
ABH03JAC4AAAcvbABB	20120624	13:27:39	8	23.7761	90.4261
ABH03JAC9AAAbWFAVV	20120624	15:27:27	51	23.7097	90.4036
ABH03JABkAAHvEkAQX	20120624	18:32:38	50	23.7386	90.4494

Table 1-2 Excerpt of the GSM data

Anonymous User ID	GSM Cell ID	Unix timestamp	Time zone
8851	712	1251762486	-7200
8851	712	1251762546	-7200
8851	836	1251762606	-7200
8851	836	1251762663	-7200

Mobile phone network records have already been extensively applied in various fields of transport such as human mobility modelling (e.g. Deville et al., 2016, Jiang et al., 2013, Isaacman et al., 2012, Song et al., 2010, Gonzalez et al., 2008), traffic model calibration (e.g. Bolla et al., 2000), origin-destination matrix estimation (e.g. Çolak et al., 2015, Iqbal et al., 2014, Pan et al., 2006, White and Wells, 2002), and trip generation estimation (e.g. Çolak et al., 2015). However, the analysis has so far been limited to extracting travel patterns and trends rather than modelling behaviour.

The main advantage with mobile phone network records is that they are already being collected on a large-scale by network operators for different purposes. For example, CDR data is collected for billing purposes, while GSM data is collected for location area updating to optimise network performance. However, the former is more widely available as it is stored for longer periods of time compared to the latter. The possibility of using these for transport studies as a by-product significantly reduces the data collection burden and cost. However, this presents significant challenges in terms of data sharing, storage, and analysis as the datasets are usually noisy, complex, large, and anonymous. In most cases, it is challenging to analyse such datasets using traditional data processing and modelling techniques as it is difficult to establish direct linkages between the observed travel patterns and the underlying causal factors, thus limiting the applicability of the data for travel behaviour and policy analysis. This motivates this research where we develop methods aimed at addressing the practical challenges associated with mobile phone network records to make them more usable for travel behaviour modelling. An important point worth noting is that the data has great potential as it is more representative, large, and frequent, thereby capturing more variability over long periods of time, which opens up a lot of scope for validation beyond what is being offered by traditional travel survey data.

The models used in this study belong to the family of discrete choice models - a well-established method for estimating econometric models of travel behaviour since in most transport choices, the options are discrete and mutually exclusive (see Ben-Akiva and Lerman, 1985 for details). However, we also note that several related studies have used machine learning as the main tool of analysis (e.g. Ellis et al., 2014, Wang et al., 2010,

Farrahi and Gatica-Perez, 2008, Sohn et al., 2006). The reason we adopt the discrete choice framework over the machine learning framework is because the former is capable of explaining the relative importance of the different factors influencing human preferences while relying on a behavioural underpinning, which is important for testing future and alternative policy scenarios, while the latter simply aim at minimising the prediction error without due consideration to established behavioural principles and interpretability (Paredes et al., 2017). Though some of the machine learning algorithms generate regression parameters, the estimates are context dependent and hardly stable, thus, it would be risky analysing these the same way we interpret typical travel behaviour model estimation outputs (Mullainathan and Spiess, 2017).

This research is timely as it presents new ways of integrating mobile phone data into traditional transport modelling approaches. As the data typically comprises of millions of records, a significant portion of this research focussed on data preparation for analysis using state-of-the-art programming languages. The research outcomes present initial efforts to provide a feasible low-cost alternative to transport planners and policy makers working in contexts with severe budget constraints on transport studies. The promising results we obtain encourage further research to improve the proposed methods, as well as extend the application of mobile phone data to other fields of travel behaviour modelling (such as mode choice). Besides, the practical challenges encountered prompted the development and/or application of novel modelling frameworks, thus the thesis makes significant methodological and applied contributions that could be useful to other fields of choice modelling.

1.2 Current progress in the application of mobile phone data to transportation studies

Prior to outlining the research objectives, it is important to review the state of the art in the application of mobile phone data to transportation studies. This review is aimed at highlighting the gaps in the literature forming the basis of the current research. The review covers seven key areas and focusses on studies using mobile phone network records (i.e. CDR and GSM data).

1.2.1 Trip generation

There have only been a few studies focusing on trip generation in the context of mobile phone data research (e.g. Çolak et al., 2015, Toole et al., 2015). These studies have only been able to produce anonymous trip rates, which cannot be used for policy assessment.

Since trip generation is mainly influenced by the trip maker's socio-demographics (see Bwambale et al., 2015 for details), such variables need to be incorporated into the models to allow for policy assessment, and previous studies have not addressed this challenge due to data anonymity.

1.2.2 Human mobility and activity patterns

This is the most widely investigated aspect of mobile phone data research. Understanding mobility and activity patterns is essential for network planning and management. However, patterns can change over time for various reasons, and most of the reviewed studies do not incorporate ways to analyse or predict the impact of such changes. The only exception is

the study by Csáji et al. (2013), who use a gravity model to explain the impact of commute distance in the extracted mobility patterns. Table 1-3 presents a summary of the reviewed studies.

1.2.3 OD matrix estimation

Related to mobility pattern generation is origin-destination (OD) matrix estimation. Mobile phone OD trips only represent a fraction of the actual OD trips due to several factors such as the mobile phone penetration rate, the market shares, and missed trips due to data noise. Usually, these trips need to be scaled to obtain an OD matrix representative of the entire population.

The issue of scaling mobile phone OD trips has been at the forefront of this subject area for many years. Goulding (2017) reviewed the current practices for calculating the OD matrix scaling factors and recommended the approach in Iqbal et al. (2014) as one of the best practices. In that approach, the scaling factors are based on the observed traffic counts at strategic locations within the road network.

However, beyond OD matrix scaling, it is important to understand the factors influencing the observed patterns. The calibration of appropriate gravity models represents effort in this direction, however, there is a need to specify more comprehensive deterrence functions, rather than those using distance only (e.g. Wang et al., 2014, Csáji et al., 2013). Table 1-4 presents an overview of the previous work in this field.

1.2.4 Mode detection

Mode choice is an important component in travel demand estimation. It influences the level of vehicular demand and therefore has a significant impact on traffic congestion. Traffic congestion itself influences various travel decisions such as route and departure time choice.

However, mode detection from anonymous mobile phone network records is very challenging due to the noisy nature of the data. This is demonstrated by the availability of few studies as summarised in Table 1-5. Since these studies have yielded promising results, there is a need for further dedicated research to improve the transferability of these approaches to real-world urban scenarios.

1.2.5 Route identification

Route choice corresponds to the final step in the four-stage model. The identification of routes from mobile phone data has been widely investigated. Table 1-6 presents an overview of the different studies in this field. It may be noted that most of the reviewed studies stop at route identification, and do not investigate the factors affecting route choice behaviour, except in Schlaich (2010), where GSM trajectories are fused with traffic state information to model the impact of variable message signs and other factors on route choice. However, GSM data is semi-continuous in nature, thus enabling the observation of full trajectories. This makes route identification relatively easy compared to when CDR data is used. This is because the latter is event-driven and more discontinuous. To date, there is no study using CDR data to analyse route choice behaviour.

1.2.6 Departure time choice

The traditional four-stage model does have some limitations. For example, it does not account for the effect of traffic congestion on the changes in departure time to avoid delays. A review of the literature shows that there is no previous study using mobile phone network records to analyse departure time choices.

The only related study is by Peer et al. (2013), who use smartphone GPS data to calculate the door-to-door travel times as part of a revealed preference departure time study, in which the desired times-of-travel are reported. Knowledge about the desired times-of-travel is critical for departure time choice modelling, however, these cannot be observed in anonymous mobile phone data. This probably explains the absence of studies in this field.

1.2.7 Population synthesis

The final transport-related application of mobile phone network records is population synthesis of mobility patterns. This technique is widely applied in activity-based modelling to generate artificial populations. The main motivation behind incorporating mobile phone data has been to generate artificial populations with realistic travel patterns. Table 1-7 presents a summary of the studies in this field.

However, in most of these studies, the disaggregate dependence structure between the user demographics and the mobile phone mobility patterns seems arbitrary. Furthermore, the assumed higher reliability of mobile phone data over travel survey data is contentious and needs to be approached impartially.

Table 1-3 Previous studies on human and activity mobility

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Candia et al. (2008)	Identifying unusual events using mobile phone data and characterising these based on methods from percolation theory.	Data type not reported	Standard percolation theory tools	Analysing the formation and decay of spatial clusters from mobile phone data can be useful in the real-time recognition of emergencies.
Farrahi and Gatica-Perez (2008)	Classifying students' daily routines according to the day of the week (weekday or weekend) and the course taken (engineering or business) by combining call log and bluetooth data.	CDR and bluetooth data	Machine learning tools (support vector machine with a gaussian kernel)	Combining proximity (bluetooth) and mobility (CDR) data can lead to significant improvements in classification when compared to using a single data source.
Choujaa and Dulay (2008)	Developing a system that relies on context sensing, time-based relationships, and anonymous landmarks for activity recognition as opposed to location data. The developed system is called TRAcME (Temporal Recognition of Activities for Mobile Environments).	GSM and bluetooth data, as well as user-reported activities for training purposes	Hierarchical learning techniques for time use graphs	The developed system can recognise multiple and consistent activities, unlike the existing techniques.
Phithakkitnukoon et al. (2010)	Investigating the correlations in the activity patterns of users based on the profiles of their work areas extracted from an activity-aware map.	CDR and point of interest (POI) data	A combination of k-means clustering and Bayes probability theory	Strong activity pattern correlations across users with the same work area profile These correlations are inversely proportional to the distance between the work areas.

Table 1-3 cont'd

Paper	Focus	Data used	Methods used	Key findings
Farrahi and Gatica-Perez (2010)	Integrating human proximity (from bluetooth data) and mobility (from CDR data) to identify human activities using probabilistic methods.	CDR and bluetooth data	Unsupervised learning techniques hinged on probabilistic topic models	Better discovery and prediction of human activity routines can be achieved using the combined data.
Yuan and Raubal (2012)	Classifying urban areas based on hourly dynamic human mobility patterns extracted from CDR data.	CDR data	Dynamic time warping algorithms	Dynamic mobility patterns can be useful in characterising different urban areas, which can help in formulating better transport and environmental policies.
Phithakkitnukoon et al. (2012)	Investigating the strength of social ties and their impact on human mobility.	CDR data	Data mining approaches and statistical analysis	Majority of the places visited were close to the nearest social tie. Geographical proximity was positively correlated with population density and inversely proportional to social tie strength.
Loibl and Peters-Anders (2012)	Investigating the mobility patterns, the population distribution, and the distribution dynamics (e.g. diurnal variations) using mobile phone data.	GSM data (location area update data)	Data mining approaches and statistical analysis	Capturing the distribution dynamics can help in the better management of demand on transportation infrastructure.

Table 1-3 cont'd

Paper	Focus	Data used	Methods used	Key findings
Yuan et al. (2012)	Investigating the relationship between mobile phone calling frequency and travel behaviour characterised by three metrics (i.e. the radius of gyration, eccentricity, and entropy).	CDR data	Data mining approaches and statistical analysis	There is a statistically significant correlation between mobile phone usage and travel behaviour.
Kang et al. (2013)	Comparing the human movements based on mobile phone data and taxicab usage in Singapore.	CDR data and GPS logs for all taxi-cabs	Data mining approaches and statistical analysis	The ratio of taxicab to mobile phone movements can help in the dynamic prediction of taxicab demand.
Jiang et al. (2013)	Generating daily mobility motifs and the purposes linked with the motif destinations using map matching techniques. The extracted motifs were compared against the trip chains obtained from travel survey data.	CDR and travel survey data	Data mining approaches, map-matching techniques, and statistical analysis	Similar trends in the shares of the daily mobility motifs extracted from both data sources, an indication that CDR data can serve as a low-cost alternative to survey data.
Paraskevopoulos et al. (2013)	Investigating and characterising usual and unusual human behaviour patterns based using call activity and mobility data.	CDR and aggregate tower-to-tower communication data	Data mining approaches, mapping tools, and statistical analysis	The observed behavioural patterns could be logically connected to major national and religious events, an indication that mobile phone data can be useful for the early detection and monitoring emergency situations.

Table 1-3 cont'd

Paper	Focus	Data used	Methods used	Key findings
Csáji et al. (2013)	Identifying home and work locations from mobile phone data and fitting a gravity model to explain the commute distances.	CDR data	Data mining approaches, mapping, statistical analysis and the gravity model	Using the number of homes and work locations improves the performance of the gravity model compared to the model using the population at the respective locations.
Calabrese et al. (2013)	Comparing the human mobility measures derived from mobile phone data (e.g. trip length) against those sourced from vehicle safety inspection data.	CDR and Odometer data	Data mining approaches, statistical analysis, and linear regression	Vehicle and mobile phone trip length have a linear positive relationship up to 65 km. Hence mobility measures from mobile phone data can be used as a proxy for estimating vehicle usage.
Kung et al. (2014)	Comparing the home-based work commute patterns across different countries (i.e. Ivory coast, Portugal, Saudi Arabia) and cities (i.e. Milan and Boston).	CDR and GPS data	Data mining approaches and statistical analysis	Commuting behaviour is generally similar across different geographical areas with minor expected differences.
Järv et al. (2014)	Exploring the longitudinal variations in human activity-travel behaviour on a monthly basis for one year.	CDR data	Data mining approaches, mapping techniques, and statistical analysis	The number of activity locations was quite stable, while the size of the activity spaces varied significantly, partly explained by seasonality and personal factors.

Table 1-3 cont'd

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Trasarti et al. (2015)	Unravelling the linkages between different regions of a city based on correlations in the temporal distributions of the corresponding population densities.	CDR data	Data mining approaches to extract the event correlation patterns for a set of regions	The study developed a "c-pattern" algorithm to help explain the logic behind the inter-regional linkages within cities, which can help in understanding the causes and effects of events.
Shi et al. (2015)	Analysing the spatial structure and characteristics of community-level human mobility using kernel density maps.	CDR data	Data mining approaches, statistical analysis, and kernel density maps	Although no spatial constraints were applied to the mobile social network, the social communities were geographically linked, an indication that individuals generally communicate with people who are near.
Jiang et al. (2017)	Generating the human mobility motifs for the entire population by combining CDR, survey, and census data using an activity-based approach.	CDR, household survey, and census data	Data mining approaches, mapping techniques, and statistical analysis	Mobile phone data can provide more realistic insights and multi-day observations of human mobility motifs compared to household survey data, which is prone to reporting errors.

Table 1-4 Previous studies on OD matrix estimation

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
White and Wells (2002)	Extracting origin-destination (OD) matrices from CDR data in a pilot study conducted by the Transport Research Laboratory (TRL) in Kent.	CDR data and an existing OD matrix for comparison	Data mining and mapping techniques	CDR data has the potential to capture realistic travel patterns with repeated observations.
Caceres et al. (2007)	A simulation-based approach to investigate the feasibility of updating an OD matrix using vehicle traffic data extracted from a GSM network.	Simulated GSM (location area update) and vehicle traffic data	Data mining approaches, simulation techniques, and statistical analysis	The automatic and somewhat real-time monitoring of mobile phone mobility provides a low-cost alternative for investigating traffic mobility.
Calabrese et al. (2011)	Developing OD matrices from mobile phone data and comparing the estimated trips with those based on census data at both the tract and county level.	CDR and census data	Data mining approaches, statistical analysis, regression, and the gravity model	Mobile phone data is a rich source of information for transport planning as it can capture the detailed spatial-temporal patterns of demand.
Iqbal et al. (2014)	Developing transient OD matrices from CDR data and scaling these using limited traffic count data to account for missed CDR trips.	CDR and video traffic count	Data mining approaches, micro-simulation, and statistical analysis	Promising results obtained, which shows that the proposed approach can serve as a low-cost option for estimating and validating travel patterns.

Table 1-4 cont'd

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Wang et al. (2014)	Optimisation of the national and regional road network using mobile phone data. Mobile phone data is used to calculate OD trips, which are then used to estimate a gravity model. The gravity model outcomes are then used for traffic assignment onto the road network, after which the impact of network changes (e.g. new road links) is tested.	CDR data	Data mining approaches, statistical analysis, mapping techniques, the gravity model, and all-or-nothing traffic assignment	Mobile phone data is able to reflect the expected behaviour in the gravity model, and can be used for network planning.
Alexander and González (2015)	Developing an OD matrix from CDR data and assigning it to the network to analyse the impact of various ride-sharing adoption rates on traffic congestion.	CDR data and various spatial and survey data sources	Data mining approaches, mapping techniques, incremental traffic assignment, and statistical analysis	Mobile phone OD matrices can be applied to successfully analyse the network-wide impact of proposed policies.
Alexander et al. (2015)	Generating OD matrices by trip purpose (i.e. HBW, HBO, and NHB) and time-of-day. The departure times are probabilistically determined using survey data from major US cities.	CDR and survey data	Data mining approaches, mapping techniques, all-or-nothing traffic assignment, and statistical analysis	CDR data can capture the travel patterns of different market segments, which is important for transport planning.

Table 1-4 cont'd

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Dong et al. (2015)	Estimating commuter OD trips from CDR data and using the estimates to divide the study area into traffic zones by applying the k-means clustering algorithm to selected semantic attributes (e.g. hourly inflow and outflow).	CDR data	Data mining approaches, k-means clustering, and statistical analysis	Mobile phone data can be used to extract valuable information on traffic flow, which can enhance our understanding of complex travel patterns.

Table 1-5 Previous studies on mode detection

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Wang et al. (2010)	Inferring transport modes (i.e. driving, public transport, and walking) from CDR data by applying k-means clustering on the travel times for each OD pair	CDR data	Data mining approaches, statistical analysis, and k-means clustering	A comparison of the travel times for each mode with those of extracted from Google maps shows that the method can reasonably detect transport modes
Doyle et al. (2011)	Using the virtual cell paths technique to extract user trajectories associated with inter-city road or rail travel from CDR data, and generating the kernel density paths for validation purposes.	CDR data and information on mode shares for validation purposes	Data mining approaches, map-matching techniques, and kernel density estimation	Kernel density maps can be used to successfully identify the routes followed and hence the travel modes used.
Qu et al. (2015)	Estimating transport mode split (car, public transport and walking) from CDR data by applying, speed-distance rules. Where the speed-distance rules do not yield a definite mode a pre-calibrated logit model is used.	CDR data, the transport network GIS data, and census data for validation purposes	Data mining approaches, logistic regression, and statistical analysis	The model correctly predicts the aggregate mode shares for the entire study area and those of a large proportion of the census tracts.

Table 1-6 Previous studies on route identification

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Schlaich et al. (2010) Schlaich (2010)	Route identification by matching the location area sequences extracted from the mobile phone data against those associated with different routes in the transport network.	GSM data	Data mining approaches and a multi-path generation algorithm	Mobile phone data can be used to continuously monitor the travel patterns for relatively long trips (i.e. trips traversing at least 3 GSM cells) due to its low spatial resolution
Saravanan et al. (2011)	Longitudinal spatial-temporal analysis of each user's CDR events to establish their daily routines and routes.	GSM data	Data mining approaches, route clustering, and statistical analysis	The method has great potential in the analysis of city-wide large-scale mobility patterns.
Görnerup (2012)	Using mobile phone data to probabilistically identify common routes using locality-sensitive hashing and graph clustering.	GSM data and GPS traces	Locality-sensitive hashing and graph clustering	The method yields promising results in terms of accurately clustering the sequences and being scalable, however, it is yet to be tested on real-life complex networks.
Tettamanti et al. (2012)	Route assignment by determining the path with the smallest sum of square distance deviations from the observed mobile phone location area sequence of a user travelling between an OD pair.	GSM data	Data mining approaches and statistical analysis	The proposed method was tried on a single OD pair and needs to be tested on a more complex network.

Table 1-6 cont'd

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Leontiadis et al. (2014)	Route searching by calculating the weights for each road segment within the cell areas linked to a user's communication events and determining the shortest weighted path for a given OD pair.	Cell phone handover and GPS data	Data mining approaches and statistical analysis	The accuracy of path determination increases with its distance potentially because individuals are more cautious of the shortest route for distant OD pairs.
Hoteit et al. (2014)	Identifying human mobility trajectories and the most crowded regions using various interpolation techniques (e.g. linear, cubic and nearest neighbour).	CDR data	Data mining approaches, mapping techniques, trajectory interpolation, and statistical analysis	A comparison of the different methods shows that cubic and linear interpolation give the best performance for commuters and inactive travellers respectively.
Nie et al. (2015)	Route identification based on the degree of similarity to a subset of k-optimal cell handover sequences extracted from the full set of possible handover sequences.	Cell phone handover data	Data mining approaches	The proposed approach yields promising results, however, it needs to be tested in a more complex network, where the number of optimal sequences may be high.

Table 1-7 Previous studies on population synthesis

Paper	**Focus**	**Data used**	**Methods used**	**Key findings**
Ros and Albertos (2016)	Updating MATSim (an agent-based multi-simulation software) by merging census and CDR data to generate synthetic populations with realistic mobility patterns.	CDR data, user demographic details (age and gender), and census data	Data mining approaches and population synthesis (Iterative Proportional Fitting)	The availability of user demographics in the CDR data ensured reliable linkages in the final synthetic population, however, such data is rarely available.
Kressner (2017)	Generating synthetic travel diaries by combining anonymous mobile phone data with traditional data sources.	CDR data, consumer data (individual-level socio-economic details), and census data	Data mining approaches and population synthesis	The aggregate validation results show that the method has great potential, however, the underlying dependency structure is unreliable.
Janzen et al. (2017)	Using population synthesis techniques to correct the underreporting of long-distance trips in travel survey data by combining the data with CDR and register data (national statistics).	CDR and register data (national statistics)	Data mining approaches and population synthesis (Iterative Proportional Fitting)	The method maintains the underlying dependency structure during population synthesis, however, the assumed uniform under-reporting in survey data and the high reliability of CDR data are unrealistic.

1.2.8 Summary of the research gaps

The previous sections have highlighted the state of the art in applying mobile phone network records to transportation studies. As described, most of these applications focus on obtaining outputs that represent the existing mobility dynamics of the study area, however, they do not incorporate mechanisms to quantify the relative importance of the different factors influencing the observed mobility trends. Without information on the underlying factors, it is difficult to conduct sensitivity analysis to predict the impact of alternative transport policies. Furthermore, the absence of policy-sensitive variables affects the ability of the models to capture the effect of disruptive changes in the study area. An example is when some links in the network are closed, which affects the overall travel time and cost between a given OD pair. The effect of such changes can only be captured if the transport models for the study area already contain such variables. Moreover, this would also improve the spatial and temporal transferability of the models as it would be easier to directly incorporate new information from the application context. As a result of the limited research effort in this direction, several other gaps have been identified in the literature.

First, although it is easy to fuse anonymous mobile phone trajectories with the attributes of the alternatives (for example distance, in the context of route choice modelling), it is more challenging to combine these trajectories with person-level attributes (for example demographic information), and previous studies have not developed data fusion frameworks to specifically address this challenge.

Secondly, there is no study showing how traditional modelling approaches (i.e. components of the 4-stage model and its extensions) can benefit from combining mobile phone data with traditional data sources (such as household surveys), except in the fields of OD matrix estimation and route choice modelling. Even under route choice modelling, there is no previous study using CDR data, which is typically easier to access, and yet presents more challenges compared to GSM data.

Thirdly, there have been limited studies systematically analysing the strengths and weaknesses of different types of mobile phone data and developing or applying appropriate frameworks to address these limitations with regard to specific modelling scenarios.

Finally, there is, to the best of our knowledge, no previous study that has attempted to use mobile phone network records to estimate the valuation metrics used in transport policy appraisal, for example, the value of travel time.

1.3 Problem statement

The fundamental role of transport models is to quantify the relative importance of the underlying factors influencing the observed travel behaviour as highlighted in the previous section. Traditionally, such models have been estimated using travel survey data, which reports the available travel options, the chosen alternatives, as well as the attributes of the alternatives and the decision maker, thereby enabling direct linkage between the observed travel behaviour and the influencing factors. However, the high cost of collecting travel survey data, coupled with stringent budget constraints on transport studies has motivated research into emerging low-cost big data sources to develop approaches that could be practically useful for transport practitioners working in contexts with severe data limitations, especially in developing countries.

With mobile phone network records offering the highest promise (due to the increasing penetration rates worldwide), most previous studies have focussed on using the data to quantify the prevailing mobility dynamics of the corresponding study areas without explaining the underlying factors influencing the observed patterns (e.g. Alexander et al., 2015, Çolak et al., 2015, Toole et al., 2015, Iqbal et al., 2014, Jiang et al., 2013, Jiang et al., 2011, Candia et al., 2008). This is due to the technical challenges associated with the data.

The arguably most significant challenge is data anonymity due to privacy reasons, which makes it difficult to perform data linkages needed to make the data usable for travel behaviour modelling. There has been limited effort to address this challenge with a few applications only seen in the field of route choice modelling (e.g. Schlaich, 2010). This thesis presents more data fusion approaches in a bid to further improve the behavioural and policy underpinnings of the resulting transport models. Moreover, in some contexts, data fusion alone may not solve the problem, especially where information on key latent indicator variables is missing (for example, the desired time of travel in the context of departure time choice modelling). In such cases, approaches to use statistical distributions to estimate these variables are proposed in this thesis.

Besides data anonymity, the other key challenges are the poor location resolution and the noisy nature of the data due to various technical reasons. These have an impact on the detection and the interpretation of the extracted trajectories. Although previous studies have discussed ways of addressing these technical limitations (e.g. Çolak et al., 2015, Iqbal et al., 2014, Jiang et al., 2013), they do not present methods to appropriately incorporate such noisy trajectories into econometric models of travel behaviour. This thesis discusses this issue and provides application examples in the different chapters.

Finally, the other significant challenge is data access and control. Mobile network operators are only willing to release datasets that do not compromise the privacy of their customers. One extreme example is the Orange data for development (D4D) dataset, where the already anonymous user IDs were scrambled across the different months to prevent possible re-identification (de Montjoye et al., 2014). Although it would have been interesting to observe the same anonymous user for the whole year, the researcher is limited to monthly observations.

While the above challenges limit the realisation of the full potential of mobile phone network records, this research hypothesises that by further probing the data coupled with innovative data fusion and modelling techniques, practical benefits can be achieved in terms of making the data more usable for travel behaviour analysis. The research focusses on demonstrating the potential of the data in the fields of trip generation, route choice, and departure time choice modelling. While doing this, the practical challenges associated with the data limited the direct application of the existing modelling frameworks. Thus the thesis makes significant methodological and applied contributions to overcome the identified limitations. The modelling scenarios investigated depended on data availability and its fitness for the purpose.

An important point worth noting is that CDR data is likely to be more useful in the near future with the increasing use of mobile internet data services (Gerpott and Thomas, 2014), which will reduce the location discontinuities in the data.

1.4 Research objectives

The main research goal is to incorporate outputs derived from mobile phone network records into traditional transport modelling approaches. This is because traditional models have the capacity to explain the behavioural relationships between phenomena, an advantage we seek to maintain. Moreover, with traditional modelling approaches, it is possible to estimate important metrics such as the value of travel time and elasticities, which are useful in transport policy appraisal. The general objectives of the research are;

G1 To develop innovative methods for combining mobile phone network records with traditional data sources to facilitate the analysis of travel behaviour;

G2 To evaluate the shortcomings of traditional modelling approaches and propose mitigation measures using mobile phone network records to optimise the reliability and applicability of the models;

G3 To analyse the limitations of mobile phone network records with regard to specific modelling scenarios and develop appropriate methods to deal with those limitations; and

G4 To assess whether models based on mobile phone network records are able to capture the expected travel behaviour in terms of the parameter estimates and/or policy insights in terms of the derived valuation metrics.

These general objectives overlap across the different chapters and are achieved by realising the following specific objectives, which largely depended on data availability and its fitness for the purpose as mentioned earlier.

S1 To develop a hybrid modelling framework fusing the CDR, GSM and demographic data of a sub-sample of users to alleviate the need for travel diary data and facilitate the subsequent analysis of trip-making behaviour using anonymous CDR data from the application context;

S2 To develop a method for modelling long-distance route choice behaviour using partial CDR trajectories combined with information from external sources;

S3 To develop a method for modelling departure time choice decisions using passively collected data without information on the desired times of travel; and

S4 To develop a method for optimising both the aggregate and the disaggregate performance of trip generation models using a combination of CDR data, household travel survey data and census data.

The framework linking the general research objectives to the specific objectives and the chapters is presented in Table 1-8.

Table 1-8 Linkages between the overall research goals and the specific objectives

	Specific objectives and the corresponding chapters			
General objectives	**S1** **Chapter 2** *Trip generation*	**S2** **Chapter 3** *Route choice*	**S3** **Chapter 4** *Departure time choice*	**S4** **Chapter 5** *Trip generation*
G1	✓	✓	✓	✓
G2	✓			✓
G3		✓	✓	
G4	✓	✓	✓	✓

1.5 Thesis outline and contributions

The subsequent chapters in this thesis (with the exception of the conclusions) correspond to papers prepared during this research. This section briefly summarises the aims of each paper and states its original contributions.

Chapter 2 presents a paper titled "Modelling trip generation using mobile phone data: a latent demographics approach". The chapter focuses on developing a novel hybrid framework that combines the CDR, GSM, and demographic data of a sub-sample of users to develop a trip generation model that mitigates the shortcomings associated with traditional models. For a sub-sample of the users, the proposed framework first estimates a demographic group prediction model based on the observed mobile phone usage behaviour (extracted from CDR data) and the reported demographics of the users. The demographic group membership probabilities from this model are then used as class weights in a latent class model for trip generation. The trip rates used for model estimation are extracted from the GSM mobility data, which is semi-continuous, and captures all the trips made by the users, thus mitigating the burden of having to fill travel diaries, which comes along with reporting errors. However, it may be noted that GSM data is typically discarded by mobile network operators as it consumes a lot of storage space. Hence the data can only be available for a sub-sample of the users during model estimation. The proposed framework overcomes this issue as it only needs anonymous CDR data during application. The data is used to extract phone usage variables, which are then used to predict the demographic group membership probabilities, and eventually the trips rates of the users. Since the CDR data for all users is stored by mobile network operators for billing purposes, it can be anonymously applied to mitigate the problems associated with the lack of detailed demographic information during model application, thus enabling more reliable prediction of aggregate travel demand. The proposed framework presents the first attempt in literature to incorporate demographic variables into models based on mobile phone data, and is an original contribution of this thesis.

Chapter 3 presents a paper titled "Modelling long-distance route choice using mobile phone call detail record data: a case study of Senegal". A review of the literature shows that most previous studies have focussed on using mobile phone data for route identification, and only a few studies based on semi-continuous GSM and GPS data have used the identified routes to analyse route choice behaviour. There is no previous study using CDR data (which is

more discontinuous and noisy) to analyse route choice behaviour. As a result of the discontinuous nature of CDR data, it is only possible to observe the partial trajectories of the users. The chapter proposes an approach for extracting and labelling the routes associated with these partial trajectories. These are then combined with travel cost, travel time, and geospatial information obtained from various sources. Due to the partial nature of the CDR trajectories, the Broad Choice Framework developed in the context of vehicle type choice (Wong, 2015) is employed to model route choice behaviour for the first time. The results show that CDR data is able to capture the behaviour towards overlapping routes, and the estimated values of travel time are reasonable. This demonstrates that CDR data can be used to inform transport policies in contexts where traditional data sources are unavailable.

Chapter 4 presents a paper titled "Modelling departure time choice using mobile phone data". Over the last few years, several studies have focussed on departure time choice modelling using GPS data, which is expensive to collect and is affected by technical issues such as signal losses and battery depletion that create gaps in the data. There has not been any study investigating the potential of mobile phone network records, which are cheaper to obtain as they are already being collected by network operators for different purposes. This chapter rigorously compares the strengths and weakness of real-world GSM and GPS data to investigate their potential use for modelling departure time decisions. It may be noted that GSM data is more appropriate for modelling departure time decisions compared to CDR data, which is event-driven and more discontinuous. The chapter presents a practical approach to impute the missing travel time information for the different departure time intervals and proposes a novel modelling framework that accounts for the fact that the desired times of travel are unobserved. The proposed framework is unique in the sense that it allows the modeller to understand the sensitivities, as well as the valuations attached to schedule delay, despite the passive nature of the data. The findings show that GSM data has fewer time gaps, which leads to more reliable model results compared to GPS data, despite the higher location accuracy of the latter. This is also supported by comparison of the valuation metrics derived from both models, where those obtained from GSM data are closer to those based on traditional data. This result could inform policy measures related to big data adoption for transportation studies.

Chapter 5 presents a paper titled "Getting the best of both worlds - a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling". This chapter describes and tests a novel joint modelling framework combining household travel survey data, census data and CDR data to optimise both the aggregate and disaggregate reliability of trip generation models. Household survey data is the most reliable source of information on travel behaviour patterns, however, the data typically covers small proportions of the population and is prone to reporting errors, which could lead to misrepresentation of the aggregate travel demand across zones. On the other hand, CDR data covers much wider proportions of the population and is a more reliable source of information on the aggregate travel patterns across zones. However, CDR data too is not error-free due to missed trips and data noise, thus the need to optimise between the household survey and the CDR data outcomes. The proposed joint modelling framework focuses on adjusting the parameter scales to optimise model performance at both the aggregate and the disaggregate level without changing the behavioural dynamics reflected in the household survey data. The results show that the proposed joint modelling framework improves both the temporal and spatial transferability of the models, thus making them more reliable. The chapter presents the first attempt in literature to combine the benefits associated

with household travel survey data and low-cost CDR data to improve the reliability of transport models, and is an original contribution of this thesis.

Chapter 6 summarises the advances made towards achieving the objectives presented in Section 1.4, linking together the different contributions, and outlining the potential directions of future research.

For purposes of clarity, the key contributions of this thesis are highlighted below in bullet points.

- Extending the application of mobile phone network records to travel behaviour modelling and policy analysis;
- A hybrid modelling framework to handle the issue of unobserved user demographics in transport models based on mobile phone data;
- A novel joint modelling framework for optimising the aggregate and disaggregate performance of models;
- Applying the broad choice framework to the context of route choice modelling using noisy CDR data; and
- A new method for modelling departure time choice without information on the desired times-of-travel.

Detailed discussions about each of these can be found in the subsequent chapters (i.e. chapters 2 to 5) and in section 6.3, where all have been summarised.

References

Alexander, L., Jiang, S., Murga, M. & González, M. C. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies,* 58**,** 240-250.

Alexander, L. P. & González, M. C. 2015. Assessing the impact of real-time ridesharing on urban traffic using mobile phone data. *Proc. UrbComp***,** 1-9.

Ben-Akiva, M. E. & Lerman, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

Bierlaire, M., Chen, J. & Newman, J. 2010. Modeling route choice behavior from smartphone GPS data.

Bolla, R., Davoli, F. & Giordano, F. Estimating road traffic parameters from mobile communications. Proceedings 7th World Congress on ITS, Turin, Italy, 2000.

Broach, J., Dill, J. & Gliebe, J. 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46**,** 1730-1740.

Bwambale, A., Choudhury, C. F. & Sanko, N. Modelling Car Trip Generation in the Developing World: The Tale of Two Cities. Transportation Research Board 94th Annual Meeting, 2015.

Caceres, N., Wideberg, J. & Benitez, F. 2007. Deriving origin–destination data from a mobile phone network. *IET Intelligent Transport Systems,* 1**,** 15-26.

Calabrese, F., Di Lorenzo, G., Liu, L. & Ratti, C. 2011. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. *IEEE Pervasive Computing,* 10**,** 36-44.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J. & Ratti, C. 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies,* 26**,** 301-313.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G. & Barabási, A.-L. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of physics A: mathematical and theoretical,* 41**,** 224015.

Choujaa, D. & Dulay, N. Tracme: Temporal activity recognition using mobile phone data. Embedded and Ubiquitous Computing, 2008. EUC'08. IEEE/IFIP International Conference on, 2008. IEEE, 119-126.

Clarke, M. 2016. Transport Sector Insight: Big Data in Transport. *The Institution of Engineering and Technology (IET)*.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Crawford, F. 2017. *Methods for analysing emerging data sources to understand variability in traveller behaviour on the road network.* University of Leeds.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z. & Blondel, V. D. 2013. Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications,* 392**,** 1459-1473.

De Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. & Blondel, V. D. 2014. D4D-Senegal: the second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*.

Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L. & Wang, D. 2016. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences***,** 201525443.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y. & Zhou, X. 2015. Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies,* 58**,** 278-291.

Doyle, J., Hung, P., Kelly, D., Mcloone, S. F. & Farrell, R. 2011. Utilising mobile phone billing records for travel mode discovery.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J. & Kerr, J. 2014. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Frontiers in public health,* 2**,** 36.

Farrahi, K. & Gatica-Perez, D. Daily routine classification from mobile phone data. International Workshop on Machine Learning for Multimodal Interaction, 2008. Springer, 173-184.

Farrahi, K. & Gatica-Perez, D. 2010. Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing,* 4**,** 746-755.

Gerpott, T. J. & Thomas, S. 2014. Empirical research on mobile Internet usage: A meta-analysis of the literature. *Telecommunications Policy,* 38**,** 291-310.

Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature,* 453**,** 779-782.

Görnerup, O. Scalable Mining of Common Routes in Mobile Communication Network Traffic Data. Pervasive, 2012. Springer, 99-106.

Goulding, J. 2017. Best Practices and Methodology for OD Matrix Creation from CDR data. United Kingdom: NLAB, University of Nottingham.

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly***,** 646-675.

GSMA 2018. The Mobile Economy 2018. London, United Kingdom: GSM Association.

Hainen, A., Wasson, J., Hubbard, S., Remias, S., Farnsworth, G. & Bullock, D. 2011. Estimating route choice and travel time reliability with field observations of Bluetooth probe vehicles. *Transportation Research Record: Journal of the Transportation Research Board***,** 43-50.

Hasan, S., Zhan, X. & Ukkusuri, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, 6.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41**,** 260-271.

Hess, S., Quddus, M., Rieser-Schüssler, N. & Daly, A. 2015. Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review,* 77**,** 29-44.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C. & Pujolle, G. 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64**,** 296-307.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40**,** 63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A. & Willinger, W. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012. Acm, 239-252.

Janzen, M., Müller, K. & Axhausen, K. W. Population Synthesis for Long-Distance Travel De-mand Simulations using Mobile Phone Data. 6th Symposium of the European Association for Research in Transportation (hEART 2017), 2017.

Järv, O., Ahas, R. & Witlox, F. 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38**,** 122-135.

Jiang, S., Ferreira, J. & González, M. C. 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data*, 3**,** 208-219.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Jiang, S., Viña-Arias, L., Zegras, C., Ferreira, J. & Gonzalez, M. Calling for Validation: Demonstrating the use of Mobile Phone data to Validate integrated land use Transportation models. International Conference Virtual City and Territory (7è: 2011: Lisboa), 2011. Department of Civil Engineering of the University of Coimbra and e-GEO, Research Center in Geography and Regional Planning of the Faculty of Social Sciences and Humanities of the Nova University of Lisbon, 181-184.

Kang, C., Sobolevsky, S., Liu, Y. & Ratti, C. Exploring human movements in Singapore: a comparative analysis based on mobile phone and taxicab usages. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, 1.

Kressner, J. D. 2017. Synthetic Household Travel Data Using Consumer and Mobile Phone Data. *Final Report for NCHRP IDEA Project 184.* Transportation Research Board.

Kung, K. S., Greco, K., Sobolevsky, S. & Ratti, C. 2014. Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9**,** e96180.

Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D. & Papagiannaki, K. From cells to streets: Estimating mobile paths with cellular-side data. Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, 2014. ACM, 121-132.

Li, L., Wang, S. & Wang, F.-Y. 2018. An Analysis of Taxi Driver's Route Choice Behavior Using the Trace Records. *IEEE Transactions on Computational Social Systems,* 5**,** 576-582.

Liu, L., Hou, A., Biderman, A., Ratti, C. & Chen, J. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference On, 2009. IEEE, 1-6.

Loibl, W. & Peters-Anders, J. 2012. Mobile phone data as source to discover spatial activity and motion patterns. *G1_Forum***,** 524-533.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F. & Liu, J. 2013. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies,* 36**,** 1-12.

Mullainathan, S. & Spiess, J. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives,* 31**,** 87-106.

Munizaga, M. A. & Palma, C. 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies,* 24**,** 9-18.

Nie, J., Zhang, J., Zhong, G. & Hu, Y. 2015. A Novel Approach to Road Matching Based on Cell Phone Handover. *CICTP 2015*.

Pan, C., Lu, J., Di, S. & Ran, B. 2006. Cellular-based data-extracting method for trip distribution. *Transportation Research Record: Journal of the Transportation Research Board***,** 33-39.

Papinski, D., Scott, D. M. & Doherty, S. T. 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation research part F: traffic psychology and behaviour,* 12**,** 347-358.

Paraskevopoulos, P., Dinh, T.-C., Dashdorj, Z., Palpanas, T. & Serafini, L. 2013. Identification and characterization of human behavior patterns from mobile phone data. *Proc. of NetMob*.

Paredes, M., Hemberg, E., O'reilly, U.-M. & Zegras, C. Machine learning or discrete choice models for car ownership demand estimation and prediction? Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on, 2017. IEEE, 780-785.

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y. & Verhoef, E. T. 2013. Door-to-door travel times in RP departure time choice models: An approximation method using GPS data. *Transportation Research Part B: Methodological,* 58**,** 134-150.

Pelletier, M.-P., Trépanier, M. & Morency, C. 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies,* 19**,** 557-568.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R. & Ratti, C. Activity-aware map: Identifying human daily activity pattern using mobile phone data. International Workshop on Human Behavior Understanding, 2010. Springer, 14-25.

Phithakkitnukoon, S., Smoreda, Z. & Olivier, P. 2012. Socio-geography of human mobility: A study using longitudinal mobile phone data. *PloS one,* 7**,** e39253.

Qu, Y., Gong, H. & Wang, P. Transportation mode split with mobile phone data. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 285-289.

Rolstad, S., Adler, J. & Rydén, A. 2011. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health,* 14**,** 1101-1108.

Ros, O. G. C. & Albertos, P. G. 2016. D5.4 Enhanced Version of MATSim: Synthetic Population Module. *Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development (INSIGHT).* Madrid, Spain: INSIGHT Consortium.

San Santoso, D. & Tsunokawa, K. 2005. Spatial transferability and updating analysis of mode choice models in developing countries. *Transportation Planning and Technology,* 28**,** 341-358.

Saravanan, M., Pravinth, S. V. & Holla, P. Route detection and mobility based clustering. Internet Multimedia Systems Architecture and Application (IMSAA), IEEE 5th International Conference, 2011. IEEE, 1-7.

Schlaich, J. 2010. Analyzing route choice behavior with mobile phone trajectories. *Transportation Research Record: Journal of the Transportation Research Board***,** 78-85.

Schlaich, J., Otterstätter, T. & Friedrich, M. Generating trajectories from mobile phone data. Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies, 2010.

Shi, L., Chi, G., Liu, X. & Liu, Y. 2015. Human mobility patterns in different communities: a mobile phone data-based social network approach. *Annals of GIS,* 21**,** 15-26.

Sohn, T., Varshavsky, A., Lamarca, A., Chen, M. Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W. G. & De Lara, E. Mobility detection using

everyday GSM traces. International Conference on Ubiquitous Computing, 2006. Springer, 212-224.

Song, C., Koren, T., Wang, P. & Barabási, A.-L. 2010. Modelling the scaling properties of human mobility. *Nature Physics,* 6**,** 818-823.

Tettamanti, T., Demeter, H. & Varga, I. 2012. Route choice estimation based on cellular signaling data. *Acta Polytechnica Hungarica,* 9**,** 207-220.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A. & González, M. C. 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*.

Trasarti, R., Olteanu-Raimond, A.-M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., Smoreda, Z. & Ziemlicki, C. 2015. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy,* 39**,** 347-362.

Tubaishat, M., Zhuang, P., Qi, Q. & Shang, Y. 2009. Wireless sensor networks in intelligent transportation systems. *Wireless communications and mobile computing,* 9**,** 287-302.

Wang, H., Calabrese, F., Di Lorenzo, G. & Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, 2010. IEEE, 318-323.

Wang, Y., Correia, G. & Romph, E. D. 2014. National and Regional Road Network Optimization for Senegal Using Mobile Phone Data. Technical report, Department of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology.

White, J. & Wells, I. Extracting Origin Destination Information from Mobile Phone Data. Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), March 2002 London. IET, pp. 30 - 34.

Wong, T. C. J. 2015. *Econometric Models in Transportation.* Ph.D. Thesis, University of California, Irvine.

Yuan, Y. & Raubal, M. Extracting dynamic urban mobility patterns from mobile phone data. International Conference on Geographic Information Science, 2012. Springer, 354-367.

Yuan, Y., Raubal, M. & Liu, Y. 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. *Computers, Environment and Urban Systems,* 36**,** 118-130.

Chapter 2
Modelling trip generation using mobile phone data: a latent demographics approach

Andrew Bwambale[*], Charisma F. Choudhury[*], Stephane Hess[*]

Abstract

Traditional approaches to trip generation modelling rely on household travel surveys which are expensive and prone to reporting errors. On the other hand, mobile phone data, where spatial-temporal trajectories of millions of users are passively recorded has recently emerged as a promising input for transport analyses. However, such data has primarily been used for the development of human mobility models, extraction of statistics on human mobility behaviour, and origin-destination matrix estimation as opposed to the development of econometric models of travel demand. This is primarily due to the exclusion of user demographics from mobile phone data made available for research (owing to privacy reasons). In this study, we address this limitation by proposing a hybrid trip generation model framework where demographic groups are treated as latent or unobserved. The proposed model first predicts the demographic group membership probabilities of individuals based on their phone usage characteristics and then uses these probabilities as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups. The model is calibrated using the call log data of a sub-sample of users with known demographics and trip rates extracted from their GSM mobility data. The performance of the hybrid model is compared with that of a traditional trip generation model which uses observed demographic variables to validate the proposed methodology. This comparative analysis shows that the model fit and the prediction results of the hybrid model are close to those of the traditional model. The research thus serves as a proof-of-concept that the mobile phone data can be successfully used to develop econometric models of transport planning by having additional information for a subset of the users.

Keywords: Trip generation, Mobile phone data, Demographic prediction

[*] Choice Modelling Centre, Institute for Transport Studies, University of Leeds (UK)

2.1 Introduction

Trip generation is the first step of the four-stage model (Ortúzar and Willumsen, 2011) and is critical to the accuracy of the subsequent stages. Generally, trip generation models establish mathematical relationships between trip making rates and the demographics of individuals or households (e.g. Bwambale et al., 2015 and the cited references). Traditional approaches to estimating trip generation models rely on household travel surveys which are expensive and prone to reporting errors. Furthermore, the application of traditional models is often hindered by the lack of detailed demographic information in the application context.

Consequently, there has been growing interest in the use of ubiquitous data for mobility modelling. Examples include social media data (e.g. Hawelka et al., 2014, Hasan et al., 2013, Wu et al., 2014), smart card data (e.g. Agard et al., 2006, Chakirov and Erath, 2012), and mobile phone data (e.g. Çolak et al., 2015, Song et al., 2010). However among these, mobile phone data has emerged as the most promising source due to the high penetration rate of mobile phones. Unique subscriber penetration in the developed world is currently very high, estimated at 79% in 2014, and projected to grow to 81% by the end of 2020, while that in the developing world was estimated at 44.6% in 2014, and is projected to grow to 56% by the end of the same period (GSMA Intelligence, 2015).

Mobile phone records, which can consist of Call Detail Records[1] (CDRs) or Global System for Mobile Communications[2] (GSM) data, have been widely used to develop human mobility models (e.g. Gonzalez et al., 2008, Jiang et al., 2013, Çolak et al., 2015, Song et al., 2010, Deville et al., 2016, Isaacman et al., 2012), calibrate traffic models (e.g. Bolla et al., 2000), develop origin-destination matrices (e.g. Iqbal et al., 2014, Pan et al., 2006, Çolak et al., 2015, White and Wells, 2002), and estimate trip rates (e.g. Çolak et al., 2015). However, they have not been used in econometric models of travel demand like trip generation, mode choice, and route choice due to missing demographic information.

The inclusion of demographic attributes into travel demand models improves their behavioural underpinning, policy sensitivity, and forecasting potential and the lack of information on such attributes is thus a valid reason for the lack of applications of mobile phone passive data in travel demand models. However, while privacy regulations make it difficult to make a 1-1 link between the socio-demographic details of a user and his/her CDRs, previous studies have demonstrated that characteristics like age, gender, employment status can be predicted by analysing the phone usage behaviour derived from the CDRs of a sub-sample of known user (e.g. Blumenstock et al., 2010, Dong et al., 2014, Brdar et al., 2012, Aarthi et al., 2011, Ying et al., 2012, Mo et al., 2012). Such techniques can be used to incorporate demographic information into human mobility models based on mobile phone data, however, there is a need to evaluate the feasibility of such an approach as this has not been tested before.

In this study, we propose a novel hybrid trip generation modelling framework to make mobile phone data usable for developing econometric models of travel behaviour and demonstrate it in the context of trip generation models. The proposed hybrid trip generation

[1] CDR data typically consists of the time stamped locations of the responding tower that handles a call/text/web access request from a user as well as the details of the request (type, sender/receiver, etc.).

[2] GSM data has more detailed location compared to CDRs and reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals.

model first predicts the demographic group membership probabilities of individuals as a function of their observed mobile phone usage. These probabilities are then used as weights inside a latent class model for trip generation, with different classes representing different socio-demographic groups.

The proposed model needs GSM locations, CDR, and the socio-demographics from a small sub-sample for estimation/calibration. However, once calibrated, it only needs anonymous CDR data to predict the trip rates. Given that CDR data is routinely saved by the mobile phone companies for billing purposes, the proposed model thus provides as a low-cost, yet reasonably accurate method for predicting trip rates – especially in the context of developing countries where traditional data is not available/reliable and acquiring large-scale GSM data is difficult (due to privacy concerns and requirement of very large storages).

We use the Nokia Mobile Data Challenge (MDC) dataset (Laurila et al., 2012, Kiukkonen et al., 2010), which is described later in this paper, to investigate the feasibility of the proposed hybrid trip generation model. We compare the goodness-of-fit of the hybrid model against that of a traditional model (which directly uses the observed demographics). We then conduct multiple runs of predictions to compare the accuracy of the trip rates predicted by the two models to validate our hypotheses that the proposed hybrid model, which only uses the predicted demographics from the CDR data, has the potential to substitute the traditional trip generation model with observed demographics.

The rest of the paper is arranged as follows. We start with a review of relevant literature, followed by an overview of the framework and the detailed model structure. We then provide a description of the data used for this study and the model estimation and validation results. Finally, we present a summary of the findings and the conclusions.

2.2 Literature review

We start by reviewing the literature on demographic prediction followed by that on passive inferring of dwell regions and trip rate extraction from mobile phone data. We end with a brief review of mathematical models of trip generation.

2.2.1 Demographic prediction from mobile phone data

The earliest attempt to use CDRs for demographic prediction was made in Rwanda (Blumenstock et al., 2010). In this study, a logit model was estimated to predict the gender of users based on the net number of calls per day and the net call duration. The study used a sample of 901 users whose demographic information was obtained through phone interviews. The estimated logit model gave a prediction accuracy of 74%. Since then, logistic regression has been applied in other demographic prediction studies (e.g. Blumenstock, 2015, Mo et al., 2012). However, most other studies have used supervised learning classification algorithms for demographic prediction. Typically, these studies involve the training of various supervised learning classifiers (e.g. Support Vector Machines and Random Forests) to make separate predictions of the demographic attributes of users based on phone usage variables (e.g. Aarthi et al., 2011, Frias-Martinez et al., 2010, Brdar et al., 2012, Ying et al., 2012, Mo et al., 2012).

Following the observation that most of the studies above had focused on predicting demographic attributes in isolation, Dong et al. (2014) investigated the possibility of improving accuracy through simultaneous demographic attribute predictions. This was

motivated by the hypothesis that mobile phone usage is influenced by a combination of demographic attributes and that separate prediction of individual demographics would reduce the probability of success due to excluded attributes. They estimated a Double Dependent Variable Factor Graph Model capable of making joint age and gender predictions based on phone usage variables and found that this improved prediction accuracy by up to 10%.

2.2.2 Inferring dwell regions from mobile phone data

In trip generation modelling, it is also important to know the home, work and other dwell regions of individuals in order to distinguish trips by purpose (e.g. Ortúzar and Willumsen, 2011). Previous studies have developed spatial-temporal algorithms for passively detecting and labelling an individual's dwell locations using CDRs (e.g. Çolak et al., 2015, Pan et al., 2006, Jiang et al., 2013, Toole et al., 2015, Akin and Sisiopiku, 2002). The nature of such algorithms depends on the accuracy used to record the location of the communication events in the CDRs. Locations are usually recorded either as triangulated mobile phone coordinates, coordinates of the cell tower that transmitted the call or as the ID of the cell from which the call was made.

Where the CDRs contain triangulated coordinates, dwell locations have been detected by applying an upper limit (usually 300m) on the distance between consecutive mobile phone coordinates and a lower limit (usually 10 minutes) on the time difference between the first and last points of a potential dwell location (e.g. Çolak et al., 2015, Jiang et al., 2013, Toole et al., 2015). For each user, the centroids of different dwell locations in close proximity to each other are then clustered into dwell regions using different clustering approaches (for example grid-based clustering (Zheng et al., 2010)) since these could be referring to the same actual point.

Where the CDRs contain cell tower coordinates, dwell locations have been detected by linking a series of consecutive communication events transmitted by cell towers in close proximity to each other followed by linking a series of consecutive events transmitted by the same tower in order to distinguish between tower jumps and actual mobile phone movements (Çolak et al., 2015). This is because mobile operators sometimes carry out tower-to-tower balancing to optimize network performance. Dwell regions are then detected by applying a lower limit (usually 10 minutes) on the time difference between the first and last records in a series of consecutive events transmitted by the same tower. A similar approach is appropriate for CDRs containing cell IDs.

Irrespective of the type of CDRs, the extracted dwell regions for each user are labelled as home, work, or other depending on the detected visitation frequency between particular times of the day, for example, home and work locations are usually defined as the most commonly visited dwell regions at night and during daytime respectively while the rest are defined as others (e.g. Çolak et al., 2015, Jiang et al., 2013). The success of the methods described above depends on the phone usage frequency of the individuals in the sample and requires long observation periods. Nevertheless, the methods can be applied to large anonymous CDR datasets to infer the dwell regions of users during trip generation model application. Methods for assigning the inferred dwell regions to Traffic Analysis Zones have been developed to ensure consistency with the existing transport models (e.g. Çolak et al., 2015, Pan et al., 2006).

2.2.3 Extraction of trip rates from mobile phone data

Previous studies have made attempts to directly estimate trip generation from CDRs (e.g. Çolak et al., 2015). CDRs have the advantage of being readily available in large quantities, however, they only report locations when the mobile phone is in use (e.g. when calls are made) making them unable to capture movements when the phone is not in use. Çolak et al. (2015) attempted to address this issue by making several assumptions e.g. by assuming an arbitrary home-based trip where the first or the last reported position of the day in the CDRs is at a non-home location. While these are reasonable assumptions, they do not properly address the issue of missed trips between communication events.

Ultimately, the best way to extract trip rates from mobile phone data is when continuous locations are provided. However, network operators usually discard such information due to its large size. Nevertheless, we note that it is feasible to store continuous location data for a reasonable sub-sample of users as was done during the Lausanne Data Collection Campaign where continuous GSM cell references were stored (Laurila et al., 2012, Kiukkonen et al., 2010).

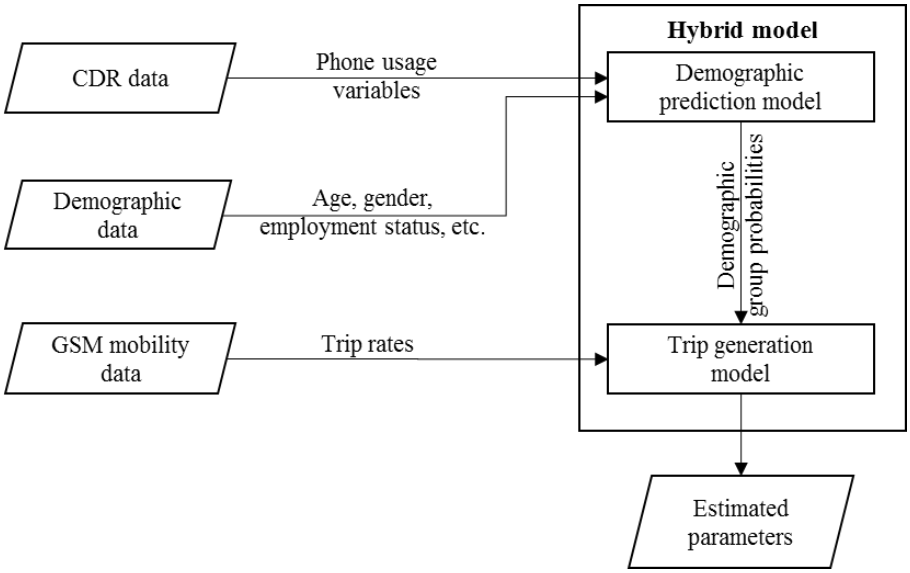
2.2.4 Mathematical models of trip generation

Discrete choice models have been the preferred approach for modelling trip generation since the ground-breaking work of McFadden (1974). This is because trip generation levels are discrete, mutually exclusive and finite. However, trip generation levels are ordered choices. An individual cannot choose to make the n^{th} trip if he/she has not previously made $(n-1)$ trips. The decision to make an additional trip depends on the number of trips already made which introduces inter-trip correlations. This has previously been taken into account using either Naturally Ordered Logit Choice Models (e.g. Vickerman and Barmby, 1985) or Ordered Response Choice Models (e.g. Bwambale et al., 2015), where the latter approach is more popular and is thus also used in this study. We also note that other trip generation modelling techniques e.g. linear regression and cross-classification (Ortúzar and Willumsen, 2011) are commonly used in practice, however, these are not considered for this study.

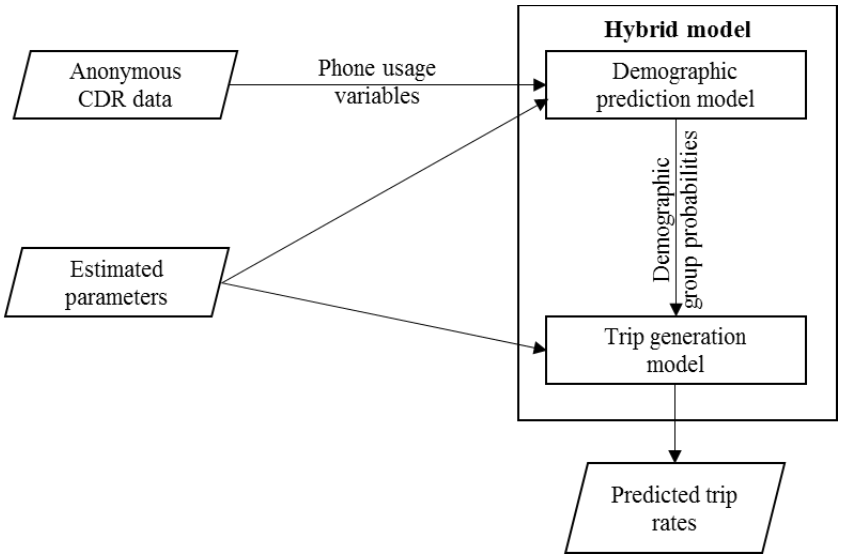
2.3 Framework

The hybrid trip generation model uses a demographic prediction model to replace the observed demographics with probabilistic latent classes of socio-demographic groups. The estimation and application frameworks are presented in Figures 2-1a and 2-1b respectively.

As presented in Figure 2-1a, the data used for estimating the hybrid trip generation model includes the GSM locations, the CDRs and the socio-demographic characteristics of a small sub-sample. The GSM data reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals and reliably captures all the trips made by the different users, except some short trips made within the boundaries of the same GSM cell. It may be noted GSM data is commonly discarded by mobile network operators due to storage space constraints and hence, though it can be stored for a small sub-sample, it is not typically available for the wider population.



a) Estimation framework (subsample of users)



b) Application framework (for all users)

Figure 2-1 Overall framework

On the other hand, CDR data, which is readily available, has a timestamped record of the phone usage activities and can be used to derive phone usage behaviour. It also records the ID of the tower that handles the call, but the location data has not been used in this case because of the availability of the GSM data which is more reliable. These data sources are used to calibrate the hybrid choice model which has two components:

1. Demographic prediction component
2. Latent class based trip generation component

In the demographic group prediction component, the demographic group membership probabilities of individuals are predicted as a function of their observed mobile phone usage (derived from CDR data). These probabilities are then used as weights inside a latent class

model for trip generation, with different classes representing different socio-demographic groups. The trip rates used for calibrating the proposed hybrid model are extracted from the GSM mobility data.

In the application stage (Figure 2-1b), the pre-estimated hybrid model uses the anonymous CDR data as the input, predicts the latent socio-demographic classes of the users using the CDR data and predicts the trip rates by feeding these latent classes to the pre-estimated trip generation model. Therefore, the socio-demographic information and the GSM data are not required in the application stage.

The detailed structure of the hybrid model is presented in Figure 2-2 and described in the following section.

2.4 Model structure

To implement the proposed hybrid framework in Figure 2-1, we propose an expanded approach that integrates two types of discrete choice mechanisms, that is, the unordered-response choice mechanism (for demographic prediction) and the ordered-response choice mechanism (for trip generation) (Greene and Hensher, 2010, Ben-Akiva and Lerman, 1985, McKelvey and Zavoina, 1975).

2.4.1 Demographic prediction

The proposed demographic prediction model is based on the Random Utility Theory (Marschak, 1960). We use the phone usage data to explain which socio-demographic group a given individual falls into. To do this, we assume that individuals in a particular demographic group are more likely to be associated with specific mobile phone usage behaviour. We use random utility theory by assuming that the segment that a given respondent falls into has the highest utility as a function of the observed phone usage behaviour.

Let U_{ns} be the utility of individual n falling in demographic group s as a function of mobile phone usage behaviour. This can be expressed as;

$$\begin{aligned} U_{ns} &= \beta_s' x_n + \xi_{ns} \\ &= \beta_s' x_n + \left(\eta' z_{ng} + \psi' h_{na} + \lambda' m_{nw} + \varepsilon_{ns}\right) \end{aligned} \quad (2\text{-}1)$$

Where x is a vector of observed phone usage variables; β_s is a vector of group-specific parameters; and ξ_{ns} is the random component of utility.

As shown, the random term comprises of the error term ε_{ns} and three demographic attribute specific components, one along each dimension of the demographic groups. η' is the gender specific constant while z_{ng} is a vector of dummy variables for the gender dimension. The additional terms ψ' and h_{na}; and λ' and m_{nw} are defined in a similar way to η' and z_{ng} but in the context of the age-group and the working status dimensions. The demographic attribute specific constants account for the unobserved phone usage dynamics that are shared across different demographic groups sharing one or more demographic attribute.

The phone usage variables are respondent specific and thus constant across the 'alternatives', which are the demographic groups. Each group has a different set of

parameters associated with it, reflecting the fact that the amount of usage has a differential impact on the likelihood of falling into a given group.

We make an assumption that the error term is independently and identically distributed across the alternatives and use the Multinomial Logit (MNL) Model (McFadden, 1974) to estimate the demographic group membership probabilities as expressed below.

$$P_{ns} = \frac{\exp\left(\beta_s' x_n + \eta' z_{ng} + \psi' h_{na} + \lambda' m_{nw}\right)}{\sum_{s^*} \exp\left(\beta_{s^*}' x_n + \eta' z_{g^*} + \psi' h_{a^*} + \lambda' m_{w^*}\right)} \quad (2\text{-}2)$$

The model parameters are then estimated by maximising the log-likelihood function below.

$$LL(\beta_s) = \sum_n \sum_s Z_{ns} \ln(P_{ns}) \quad (2\text{-}3)$$

Where $Z_{ns} = 1$ if and only if individual n belongs to demographic group s otherwise, $Z_{ns} = 0$.

As a result, each respondent has a non-zero probability of falling into each of the different socio-demographic groups, but the more the model is able to link the socio-demographic characteristics to phone usage, the more deterministic the allocation to these groups becomes in the model.

2.4.2 Trip generation

As mentioned, the ordered response choice mechanism is used in the trip generation model component considering the ordered nature of trip generation choices. This mechanism assumes that every individual has a latent continuous trip making propensity that is a function of his/her demographics, which is then converted to discrete trips using estimated cut-off points (Greene and Hensher, 2010, McKelvey and Zavoina, 1975). We first present the traditional framework (where demographics are observed) and then present our proposed extension that addresses the issue of unobserved demographics.

2.4.2.1 The traditional trip generation model (with observed demographics)

Let h_n^* be the latent trip-making propensity for individual n based on his observed demographic attributes. Using the ordered-response choice mechanism, this can be expressed as;

$$h_n^* = \gamma' w_n + \varepsilon_n \quad (2\text{-}4)$$

$$t = \begin{cases} < 10, & if\ h_n^* \leq \delta_0 \\ 10 - 15, & if\ \delta_0 < h_n^* \leq \delta_1 \\ 16 - 20, & if\ \delta_1 < h_n^* \leq \delta_2 \\ 21 - 25, & if\ \delta_2 < h_n^* \leq \delta_3 \\ > 10, & if\ h_n^* > \delta_3 \end{cases}$$

Where; w_n is a vector of the observed demographic attributes for individual n ; ε_n is the random error term; γ' is a vector of the model coefficients; t is the number of trips per week; and $\delta_0 < \delta_1 < \delta_2 < \delta_3$ are the cut-off points. Note that different categorisations of the weekly trip rates were tested and these were found to provide the best model fit. We make

an assumption that the random error term follows a Gumbel Distribution and use the Ordered Response Logit (ORL) Model (Greene and Hensher, 2010, McKelvey and Zavoina, 1975) to estimate the trip generation probabilities as expressed below;

$$\begin{aligned}
&P_{n,\ t<10} = \Lambda(\delta_0 - \gamma' w_n) \\
&P_{n,\ 10\le t\le 15} = \Lambda(\delta_1 - \gamma' w_n) - \Lambda(\delta_0 - \gamma' w_n) \\
&P_{n,\ 16\le t\le 20} = \Lambda(\delta_2 - \gamma' w_n) - \Lambda(\delta_1 - \gamma' w_n) \\
&P_{n,\ 21\le t\le 25} = \Lambda(\delta_3 - \gamma' w_n) - \Lambda(\delta_2 - \gamma' w_n) \\
&P_{n,\ t>25} = 1 - \Lambda(\delta_3 - \gamma' w_n)
\end{aligned} \tag{2-5}$$

Where $\Lambda(q) = \exp[-\exp(-q)]$ represents the standard cumulative Gumbel Distribution. The model parameters are then estimated by maximising the log-likelihood function below.

$$LL(\gamma, \delta) = \sum_n \sum_t Z_{nt} \ln(P_{nt}) \tag{2-6}$$

2.4.2.2 The hybrid trip generation model (with predicted demographics)

Let ${y_{n|s}}^*$ be the latent trip-making propensity for individual n on condition that he/she belongs to latent demographic group s. This latent propensity can be expressed as a function of the typical demographic attributes associated with latent demographic group s as shown below;

$${y_{n|s}}^* = \alpha' w_{n|s} + \varepsilon_{n|s} \tag{2-7}$$

$$t = \begin{cases} < 10, & if\ {y_{n|s}}^* \le \mu_0 \\ 10-15, & if\ \mu_0 < {y_{n|s}}^* \le \mu_1 \\ 16-20, & if\ \mu_1 < {y_{n|s}}^* \le \mu_2 \\ 21-25, & if\ \mu_2 < {y_{n|s}}^* \le \mu_3 \\ > 10, & if\ {y_{n|s}}^* > \mu_3 \end{cases}$$

Where; $w_{n|s}$ is a vector of the typical demographic attributes for individual n given that he/she is associated with latent demographic group s; $\varepsilon_{n|s}$ is the random error term; t is the number of trips; $\mu_0 < \mu_1 < \mu_2 < \mu_3$ are the cut-off points; and α' is a vector of the model coefficients.

We again assume that the random error term follows a Gumbel Distribution and estimate the conditional trip generation probabilities as expressed below;

$$\begin{aligned}
&P_{n,\ (t<10)|s} = \Lambda(\mu_0 - \alpha' w_{n|s}) \\
&P_{n,\ (10\le t\le 15)|s} = \Lambda(\mu_1 - \alpha' w_{n|s}) - \Lambda(\mu_0 - \alpha' w_{n|s}) \\
&P_{n,\ (16\le t\le 20)|s} = \Lambda(\mu_2 - \alpha' w_{n|s}) - \Lambda(\mu_1 - \alpha' w_{n|s}) \\
&P_{n,\ (21\le t\le 25)|s} = \Lambda(\mu_3 - \alpha' w_{n|s}) - \Lambda(\mu_2 - \alpha' w_{n|s}) \\
&P_{n,\ (t>25)|s} = 1 - \Lambda(\mu_3 - \alpha' w_{n|s})
\end{aligned} \tag{2-8}$$

These calculations are conditional on knowing the socio-demographics of respondent n, reflected by that respondent falling into demographic group s. However in reality, we do

not know which class the respondent falls into. Therefore, the unconditional trip generation probabilities are estimated as the weighted averages of the conditional probabilities as expressed in Equation 2-9. The weights P_{ns} are the demographic group membership probabilities estimated from Equation 2-2 at the maximum likelihood estimates.

$$
\begin{aligned}
P_{n,\ t<10} &= \sum_{s} P_{ns} * \Lambda(\mu_0 - \alpha' w_{n|s}) \\
P_{n,\ 10\le t\le 15} &= \sum_{s} P_{ns} * \left[\Lambda(\mu_1 - \alpha' w_{n|s}) - \Lambda(\mu_0 - \alpha' w_{n|s})\right] \\
P_{n,\ 16\le t\le 20} &= \sum_{s} P_{ns} * \left[\Lambda(\mu_2 - \alpha' w_{n|s}) - \Lambda(\mu_1 - \alpha' w_{n|s})\right] \\
P_{n,\ 21\le t\le 25} &= \sum_{s} P_{ns} * \left[\Lambda(\mu_3 - \alpha' w_{n|s}) - \Lambda(\mu_2 - \alpha' w_{n|s})\right] \\
P_{n,\ t>25} &= \sum_{s} P_{ns} * \left[1 - \Lambda(\mu_3 - \alpha' w_{n|s})\right]
\end{aligned}
\qquad (2\text{-}9)
$$

The advantage with ordered response models is their parsimonious structure as they are specified using monotonic parameters (see the γ and α parameters in Equations 2-4 and 2-7 respectively). This simplified model structure is based on the hypothesis that the data supports the proportional odds assumption, where the logarithms of the cumulative odds of the ordered alternatives increase proportionally (Greene and Hensher, 2010, McCullagh, 1980). If the data does not support this assumption, the model could lead to biased parameter estimates.

To test whether the proportional odds assumption holds, Borooah (2001) proposes a likelihood ratio test where the ordered response model (with monotonic parameters) is compared against an equivalent MNL model (with alternative-specific variable parameters), and the calculated likelihood ratio is assessed with respect to a chi-square statistic where the degrees of freedom correspond to the difference in the number of parameters. A statistically insignificant result suggests that there is no cause for concern. It may be noted that this was the case in this study where the p-values of the chi-square statistics were 0.8968 and 0.8153 for the traditional and the hybrid models respectively.

Figure 2-2 presents the full path diagram of the hybrid model structure where unobserved variables are shown in circles and observed variables in rectangles. It may be noted that two sequential estimators are used to estimate the hybrid framework. The first results in parameters that provide the best fit for the demographic prediction model and the second results in parameters that provide the best fit for the trip generation model. This is different from a simultaneous estimator which tries to jointly predict both the demographic groups and the trip generation levels. A simultaneous model could lead to gains in efficiency (i.e. smaller standard errors) but also opens up risk in terms of confounding between the two model components.

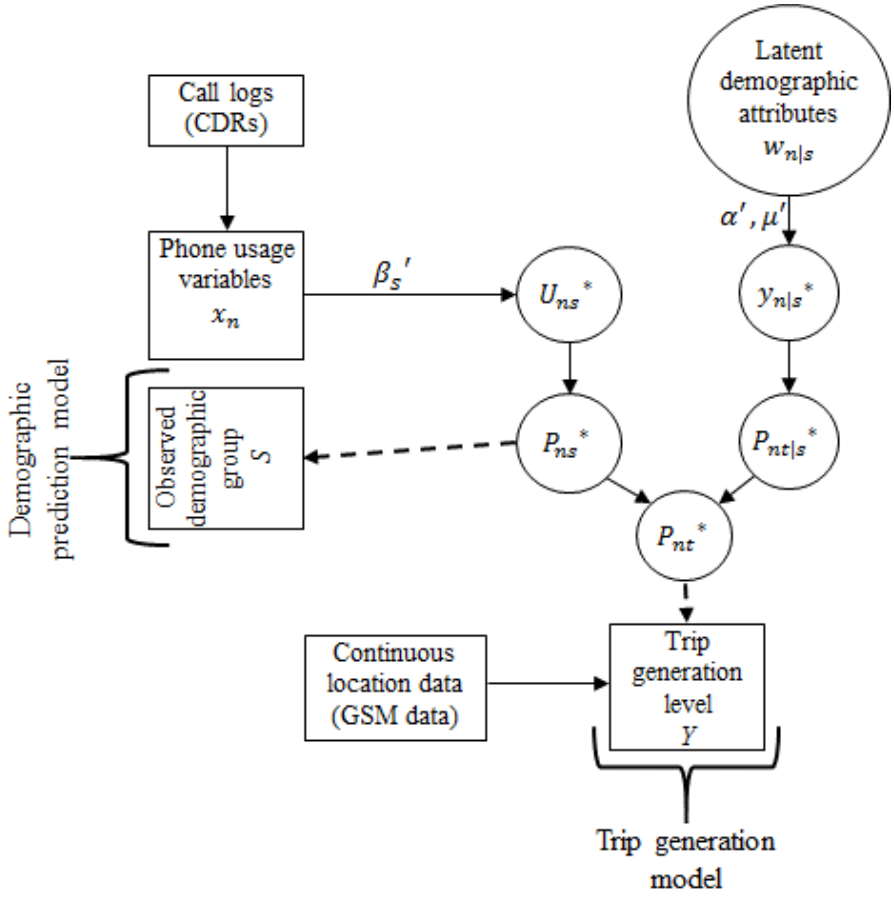


Figure 2-2 Full path diagram of the hybrid model structure

Notation

α, β, μ	Vectors of unknown parameters to be estimated
X	Phone usage variables
S	Demographic groups
t	Number of trips
$U_{ns}{}^*$	Utility of individual n falling in demographic group s
$P_{ns}{}^*$	Membership probability to demographic group s for individual n
$w_{n\|s}$	A vector of the typical demographic attributes for individual n given that he is associated with latent demographic group s
$y_{n\|s}{}^*$	Latent trip-making propensity for individual n on condition that he belongs to latent demographic group is s
$P_{nt\|s}{}^*$	Conditional probability for making t trips given that the latent demographic group for individual n is s
$P_{nt}{}^*$	Unconditional probability for making t trips for individual n
Y	Number of trips made

2.4.3 Evaluation criteria for model performance

For model evaluation, we compare the goodness-of-fit during estimation and validation. For the estimation, we use the adjusted-rho square and the likelihood ratio test (Ben-Akiva and Lerman, 1985) which are defined as follows, respectively;

$$\rho_{adj}{}^2 = 1 - \frac{LL(F) - k}{LL(0)} \quad \text{and} \quad LR = -2[LL(0) - LL(F)] \qquad (2\text{-}10)$$

Where; k is the number of model parameters, $LL(F)$ and $LL(0)$ are the values of the log-likelihood function at convergence and at zero respectively.

For model validation, a hold-out sample (not used for model estimation) is used to confirm that the estimation results are not simply due to overfitting. In this stage, we use both aggregate and disaggregate measures of fit. At the aggregate level, we compare the predicted and actual shares and compute the Root Mean Square Error (RMSE). At the disaggregate level, we use the predictive rho-square and the average probability of correct prediction. The predictive rho-square is obtained by calculating the log-likelihood for the validation sample at the pre-estimated maximum likelihood parameters and at zero and then applying Equation 2-10 without the k. The average probability of correct prediction is obtained by computing the mean probability of success for the validation sample based on the pre-estimated maximum likelihood parameters.

2.5 Data

We use data from the Nokia Mobile Data Challenge (MDC) for this study (Laurila et al., 2012, Kiukkonen et al., 2010). The data was generated during the Lausanne Data Collection Campaign by 158 participants with known demographics. These participated in the campaign at different time periods between 2010 and 2012, each lasting several months. This makes the data rich in terms of temporal coverage. The full database contains several types of smartphone records (e.g. Bluetooth usage data), however, we only use the call logs and the GSM cells data (mobility data) to improve the transferability of our approach. The subsequent sections briefly describe the data used including the analysis undertaken.

2.5.1 Extraction of demographic groups from the demographic data

The demographic data file contains the demographics of 158 participants. Each record in this file is described by; a user ID, the gender, the age-group, and the working status of the participant, among others (e.g. marital status). Out of these, 4 participants were disregarded because they had missing demographic information, leaving 154 participants. Demographic groups were formed by generating various possible combinations of age-group, gender, and working status. In total, seven demographic groups were observed in the data as shown in Table 2-1, where some of the demographic groups have very small sub-samples. This problem could have been avoided by conducting demographically stratified random sampling of the participants in the data collection phase (which was beyond our control).

2.5.2 Extraction of phone usage variables from the call log data

The call log data file contains a register of all the communication events of the participants (calls and short messages). In total, there are over 0.42 million call log events. Each call log

event is described by; a user ID, the time of the call, the status of sent short messages, the direction of the call, the type of call, the other party's anonymized phone number, and the call duration. The information in this file is equivalent to what would be found in CDRs. The call log data was analysed to extract several phone usage variables based on guidance from previous literature (e.g. Aarthi et al., 2011, Blumenstock, 2015, Blumenstock et al., 2010, Frias-Martinez et al., 2010) and intuition. Table 2-1 presents the summary statistics of the extracted phone usage variables.

Table 2-1 Summary statistics

Demographic group summary statistics

Demographic group	Assigned code	Number of participants	Percentage, %
Female non-worker aged below 21 years	F-NW-U21	7	4.9
Female worker aged above 21 years	F-WO-A21	29	20.3
Female non-worker aged above 21 years	F-NW-A21	19	13.3
Male non-worker aged below 21 years	M-NW-U21	4	2.8
Male worker aged below 21 years	M-WO-U21	3	2.1
Male non-worker aged above 21 years	M-NW-A21	22	15.4
Male worker aged above 21 years	M-WO-A21	59	41.3
Total		**143**	**100**

Sample phone usage summary statistics (extracted from call log data)

Variable	Statistic
Average number of outgoing calls per user, per day	3.4
Average number of incoming calls per user, per day	1.5
Average number of outgoing short messages per user, per day	1.7
Average number of incoming short messages per user, per day	2.5
Average number of missed calls per user, per day	0.7

Trip generation summary statistics (extracted from GSM data)

Number of trips per week from home	Number of participants	Percentage, %
< 10	4	2.8
10 – 15	26	18.2
16 – 20	43	30.1
21 – 25	14	9.8
> 25	56	39.2
Total	**143**	**100**

2.5.3 Extraction of trip rates from the GSM (mobility) data

The GSM data file contains a register of all the GSM cells seen by the participants' mobile phones at an interval of approximately 60 seconds. This data file contains over 50.8 million records generated by all the participants. Each GSM record is described by; a user ID, a unique internal ID for the GSM cell, and the record creation time and date. The GSM data was analysed to extract the number of trip origins from home using the following approach.

First, all the GSM cell IDs seen at night (between 8 pm and 6 am) by the different user IDs were extracted and ordered according to the record creation time and date. For each date, the GSM cell ID seen for the longest continuous time at night was established and the most

common among these across the different dates determined as the home GSM cell for the user ID. The weekly trip rates from home were then estimated by analysing the GSM mobility data to determine the number of times per week the different user IDs were not seen in their respective home GSM cells for periods longer than 10 minutes. We considered 10 minutes as the appropriate threshold for distinguishing between actual trips and tower jumps. We do not classify the trips by purpose because the geographical locations of the GSM cells have been anonymised thereby making it difficult to infer activities by map matching. We acknowledge that the resolution of the GSM mobility data only enables the capture of trips made outside the home GSM cell and misses short trips made within the boundaries of the home GSM cell. Our approach is therefore suitable for urban areas such as Lausanne where GSM cell sizes can be as small as 100m (De Groote, 2005). At this stage, we disregarded 11 participants who had incomplete weeks of data, leaving 143 participants. Table 2-1 presents the summary statistics of the extracted trip rates.

2.6 Estimation results

In this section, we present the estimation results for both the demographic group prediction model and the trip generation models based on the full sample.

2.6.1 Demographic group prediction model

Table 2-2 presents the estimation results of the demographic prediction model. We tested various combinations of phone usage variables in terms of the statistical performance of the associated parameters and the overall model performance and settled for a set of eleven shown in Table 2-2. We found that differentiation of phone usage by time segment (e.g. working hours and night) was statistically important for most of the variables while differentiation by weekdays versus weekends was not. We also found that interacting some of the variables (e.g. net = outgoing - incoming) was statistically important, however, we acknowledge that we have not exhausted all the possibilities.

The parameters of the demographic prediction model represent the effect of the variables on the utility of each demographic group relative to that of the reference group M-WO-A21 (male workers aged above 21 years). We do not have a priori expectations of the parameter signs since this is still a new area of research, moreover, mobile phone usage behaviour is likely to differ from place to place. Therefore, we analyse this particular case using our intuitive reasoning. To do this, we first analyse the demographic attribute specific constants. Among these, we find that the only statistically significant constant is that associated with individuals above 21 years. This indicates the existence of statistically strong unobserved phone usage dynamics common across different demographic groups sharing the same age-group. The rest of the constants are statistically insignificant probably because the associated phone usage dynamics have been captured by the specified explanatory variables.

We then analyse the parameters of the demographic groups having only one attribute not in the reference group M-WO-A21 so as to establish the unique effect of each attribute. These groups (and the complement attributes) are; F-WO-A21 (female), M-WO-U21 (age below 21 years), and M-NW-A21 (non-worker). See Table 2-1 for the group definitions.

Table 2-2 Parameter estimates of the demographic prediction model
(See Table 2-1 for the parameter definitions)

Variable	Parameter	t-statistic
Net number of calls (outgoing – incoming) in the morning (06:00 AM – 08:00 AM)		
F-NW-U21	-2.9394	-0.82
F-WO-A21	-3.6428	-1.92
F-NW-A21	-0.9796	-0.32
M-NW-U21	-40.3761	-3.69
M-NW-A21	-0.6865	-0.32
M-WO-U21	-18.6114	-3.77
Number of outgoing calls at lunch time (01:00 PM – 02:00 PM)		
F-NW-U21	-7.0073	-1.94
F-WO-A21	-9.6277	-2.91
F-NW-A21	1.2208	0.60
M-NW-U21	-6.0815	-1.47
M-NW-A21	-0.2364	-0.10
M-WO-U21	4.5906	1.90
Net number of calls (outgoing – incoming) during working hours (08:00 AM – 01:00 PM and 02:00 PM – 05:00 PM)		
F-NW-U21	1.9960	3.15
F-WO-A21	2.4266	3.30
F-NW-A21	-0.2648	-0.34
M-NW-U21	1.9878	1.99
M-NW-A21	-0.4622	-0.64
M-WO-U21	2.0136	3.61
Number of outgoing calls in the evening (05:00 PM – 08:00 PM)		
F-NW-U21	-1.3946	-0.99
F-WO-A21	0.5150	0.43
F-NW-A21	1.4971	1.61
M-NW-U21	0.8424	0.26
M-NW-A21	2.4243	2.54
M-WO-U21	0.3342	0.24
Net number of calls (outgoing – incoming) at night (08:00 PM – 06:00 AM)		
F-NW-U21	-0.2181	-0.11
F-WO-A21	-4.8651	-1.99
F-NW-A21	-2.3378	-1.01
M-NW-U21	2.4107	0.56
M-NW-A21	0.7189	0.31
M-WO-U21	0.9151	0.31
Number of outgoing short messages during working hours (08:00 AM – 01:00 PM and 02:00 PM – 05:00 PM)		
F-NW-U21	1.6257	2.64
F-WO-A21	0.7478	1.21
F-NW-A21	0.0456	0.08
M-NW-U21	1.9497	2.7
M-NW-A21	1.2729	1.47
M-WO-U21	-1.4161	-1.97

Table 2-2 cont'd

Variable	Parameter	t-statistic
Number of outgoing short messages at night (08:00 PM – 06:00 AM)		
F-NW-U21	-1.5717	-1.65
F-WO-A21	-1.2529	-0.99
F-NW-A21	0.5355	0.61
M-NW-U21	-1.166	-1.04
M-NW-A21	-3.3759	-1.70
M-WO-U21	2.1387	2.30
Average duration of outgoing calls in the evening (05:00 PM – 08:00 PM)		
F-NW-U21	0.0032	0.93
F-WO-A21	0.0020	0.76
F-NW-A21	0.0001	0.05
M-NW-U21	0.0516	3.04
M-NW-A21	0.0005	0.16
M-WO-U21	-0.0028	-0.94
Average duration of incoming calls in the evening (05:00 PM – 08:00 PM)		
F-NW-U21	0.0017	1.08
F-WO-A21	0.0008	0.57
F-NW-A21	0.0013	1.01
M-NW-U21	-0.0897	-2.94
M-NW-A21	0.0002	0.15
M-WO-U21	0.0018	0.86
Outdegree of the social network		
F-NW-U21	-0.0360	-1.89
F-WO-A21	-0.0063	-0.72
F-NW-A21	0.0016	0.18
M-NW-U21	0.1579	2.24
M-NW-A21	-0.0076	-0.68
M-WO-U21	0.0083	0.57
Indegree of the social network		
F-NW-U21	0.0238	1.25
F-WO-A21	0.0037	0.45
F-NW-A21	-0.0163	-1.48
M-NW-U21	-0.1633	-2.00
M-NW-A21	-0.0088	-0.78
M-WO-U21	-0.0357	-1.62
Demographic attribute specific constants		
Males	0.0401	0.08
Workers	-0.0205	-0.03
Individuals > 21 years	2.1583	1.86
Measures of fit		
Number of observations	143	
Log-likelihood at zero	-278.27	
Log-likelihood at convergence	-169.83	
Number of parameters	69	
Adjusted-rho square	0.14	
Likelihood ratio	216.9	
Chi-square statistic (69, 0.05)	89.39	

From Table 2-2, it is observed that the net number of calls and outgoing short messages during working hours, the number of outgoing calls and the total call duration (outgoing and

incoming) in the evening, and the social network indegree (the unique number of incoming contacts) have positive parameter signs for the F-WO-A21 group. However among these, the only statistically significant parameter is that for the net number of calls during working hours. This suggests that females in comparison to males tend to use their phones more during working hours. On the other hand, the net number of calls in the morning and at night, the number of outgoing short messages at night, the number of outgoing calls during lunch time, and the social network outdegree (the unique number of outgoing contacts) have negative parameter signs for the same group. Most of these parameters are statistically significant except those for the outgoing short messages at night, and the social network outdegree. This suggests that males in comparison to females tend to make more phone calls during non-working hours since majority of them are workers.

Similarly, it is observed that the number of outgoing calls and the total call duration (outgoing and incoming) in the evening, the net number of calls at night, and the number of outgoing short messages during working hours have positive parameter signs for the M-NW-A21 group. However, the only statistically significant parameter that for the number of outgoing calls in the evening. This points to the idea that workers in comparison to non-workers tend to call fewer people in the evenings probably because they prefer to utilize this time preparing for the next day.

On the other hand, the net number of calls in the morning and during working hours, the number of outgoing calls at lunch, the number of outgoing short messages at night, and the social network indegree and outdegree have negative parameter signs for the M-NW-A21 group. However, the only statistically significant parameter (at the 90% confidence level) is that for the number of outgoing short messages at night. This implies that workers in comparison to non-workers tend to send out more short messages at night probably because they do not have time to do so during the day due to work.

In addition, it is observed that the number of outgoing calls during lunch time and in the evening, the net number of calls during working hours and at night, the number of outgoing short messages at night, the average duration of incoming calls in the evening, and the social network outdegree have positive parameter signs for the M-WO-U21 group. However, the only statistically significant parameters among these are those for the net number of calls during working hours, the number of outgoing calls during lunch time, and the number of outgoing short messages at night. This is a reflection of the possibility that individuals aged below 21 years tend to make more phone calls while at work and during lunch breaks and also send more short messages at night. On the other hand, the net number of calls in the morning, the number of outgoing short messages during working hours, the average duration of outgoing calls in the evening, and the social network indegree have negative parameter signs for the same group. However, the only parameters that are statistically significant among these are those for the net number of calls in the morning and the number of outgoing short messages during working hours. This indicates that individuals aged below 21 years in comparison to those above 21 years tend to make fewer calls in the morning and send fewer short messages during working hours (because they already make more calls during this time as earlier noted).

The last rows of Table 2-2 provide the measures of fit in estimation. From these, it is noted that the model passes the likelihood ratio test at the 95% confidence level in comparison

with a model giving equal probabilities to the different groups for each individual (Ben-Akiva and Lerman, 1985).

2.6.2 Trip generation models

The variables commonly used in trip generation models include; income, car ownership, working status, age, and gender (e.g. Bwambale et al., 2015). Among these, income and car ownership were not available in the MDC dataset and hence could not be considered in the demographic prediction model, therefore, we only considered gender, working status, and age. The estimation results of the hybrid model (which uses the predicted demographics) are presented in Table 2-3 alongside those of a model which uses the observed demographics (referred as the traditional model in the following sections).

Table 2-3 Parameter estimates of the trip generation models

Variable	**Traditional model** *(With observed demographics)*		**Hybrid model** *(With predicted demographics)*		**t-statistic of the difference in parameters**
	Parameter	**t-statistic**	**Parameter**	**t-statistic**	
Dummies specific to					
Males	-0.3981	-2.33	-0.9125	-2.90	-1.43
Workers	-0.3962	-2.24	-0.7817	-1.65	-0.76
Individuals > 21 years	0.6020	1.91	1.0127	1.80	0.64
Cut-off points specific to					
Trips per week <10	-1.2775	-4.19	-1.6997	-3.82	-0.78
Trips per week 10 - 15	-0.4221	-1.35	-0.7087	-1.69	-0.55
Trips per week 16 – 20	0.4647	1.44	0.3119	0.73	-0.29
Trips per week 21 - 25	0.7726	2.34	0.6542	1.49	-0.22
Measures of fit in the estimation sample					
No. of observations	143		143		
Zero log-likelihood	-230.2		-230.2		
Sample shares log-likelihood	-197.07		-197.07		
Final log-likelihood	-191.8		-192.3		
Number of parameters	3		3		
Adjusted-rho square	0.14		0.13		
Likelihood ratio w.r.t sample shares	10.54		9.54		
Chi-square stat (3, 0.05)	7.81		7.81		

The parameters of these models indicate the effect of the variables on the trip making propensity of individuals. The signs of all the parameters are consistent with a priori expectations. When individuals are employed, they are always engaged at the workplace and tend to travel less frequently to and from home in comparison to non-workers, hence the negative parameter sign for workers. Similarly, females generally run more errands (e.g. shopping, taking children to school etc.) irrespective of who pays the costs. Therefore, females tend to travel more frequently to and from home in comparison to males, hence the negative parameter sign for males. On the other hand, individuals above 21 years are generally out of school and if not already employed, are usually in active search for employment opportunities which requires a lot of travel hence the positive parameter sign.

To further interpret the estimated parameters, it would have been necessary to compute the odd ratios with respect to unit (0,1) changes in each of the demographic variables while keeping the others fixed, however, this was difficult to achieve for the hybrid model, which relies on demographic group membership probabilities that represent the joint likelihood of more than one demographic variable (i.e. age, gender, and working status), and the probabilities cannot shift from 0 to 1 (see Equation 2-9). Nevertheless, the signs of all the parameters for both models are the same and the t-statistics for the differences between the parameters are insignificant. This shows that both models capture the same trip generation behaviour and would lead to similar policy conclusions.

The last rows of Table 2-3 provide the measures of fit in estimation. The adjusted-rho square values and the final log-likelihoods show that the traditional model performs slightly better than the hybrid model. This is to be expected given the error-free measures of the socio-demographics used in the traditional model. On the other hand, it is also worth acknowledging that part of the performance of the hybrid model could be due to allowing for heterogeneity through the probabilistic component as individuals are not assigned deterministically to classes.

2.7 Validation results

In order to compare the predictive power of the traditional and the proposed hybrid model, we randomly split the data into five parts at the person level and generated five rolling subsets, each comprising of 80% of the data for model estimation purposes. For each of the five estimation subsets, we generated a complementary subset comprising of 20% of the data for validation purposes.

We estimated models based on each of the five estimation subsets and the general interpretation of the model results remains the same. Table 2-4 presents the measures of fit of the models based on each of the subsets. As can be observed, the final log-likelihoods of the hybrid model remain close to those of the traditional model across the different subsets of the data. We tested the predictive power of each of these models using the corresponding complementary subsets. The subsequent sections present the validation results of both the demographic group prediction model and the trip generation models.

Table 2-4 Final log-likelihoods of the models on the estimation subsets

Model	**Log-likelihood**	**Subset 1** *N=115*	**Subset 2** *N=114*	**Subset 3** *N=114*	**Subset 4** *N=114*	**Subset 5** *N=115*
Demographic prediction model	Initial	-223.78	-221.83	-221.83	-221.83	-223.78
	Final	-121.21	-130.74	-135.37	-120.48	-122.64
Hybrid trip generation model *(With predicted demographics)*	Initial	-185.09	-183.48	-183.48	-183.48	-185.09
	Final	-153.66	-151.90	-155.14	-155.32	-154.68
Traditional trip generation model *(With observed demographics)*	Initial	-185.09	-183.48	-183.48	-183.48	-185.09
	Final	-152.34	-151.23	-154.50	-154.14	-152.30

2.7.1 Demographic group prediction

We start by assessing the predictive performance of the demographic prediction model using the five validation subsets. The actual and predicted demographic group shares in the validation subsets are presented in Figure 2-3.

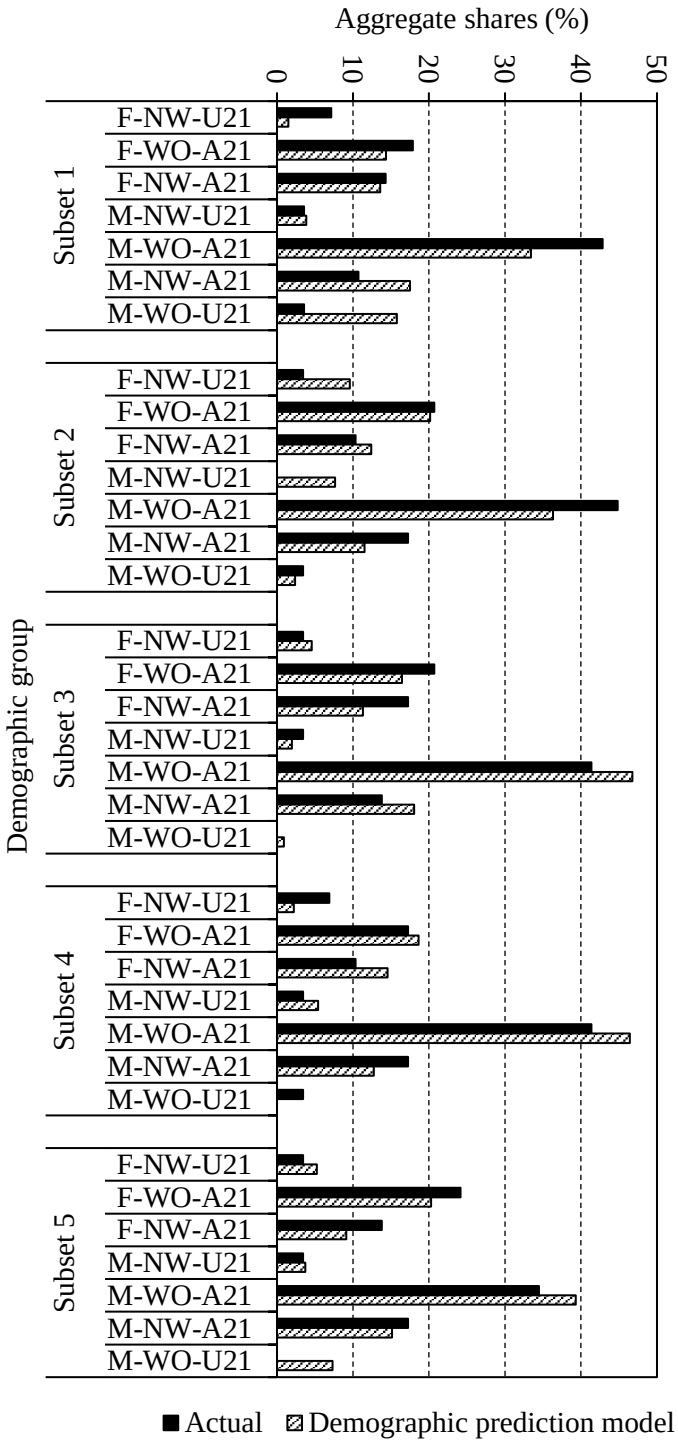


Figure 2-3 Demographic model predictive performance in the validation subsets
(See Table 2-1 for the demographic group definitions)

As can be observed, both the actual and predicted shares tend to follow a similar trend albeit with observable differences across all the five subsets. This is probably due to weaknesses in variable specification, however, there is a possibility that more sophisticated models (e.g. the mixed logit model) could improve the performance. Nevertheless, the similarity in trends is a good starting point and motivates further research to improve our approach.

2.7.2 Trip generation

In this section, we assess the predictive performance of both the traditional and the hybrid trip generation models using the five validation subsets. The actual and predicted trip generation shares in the validation subsets are presented in Figure 2-4.

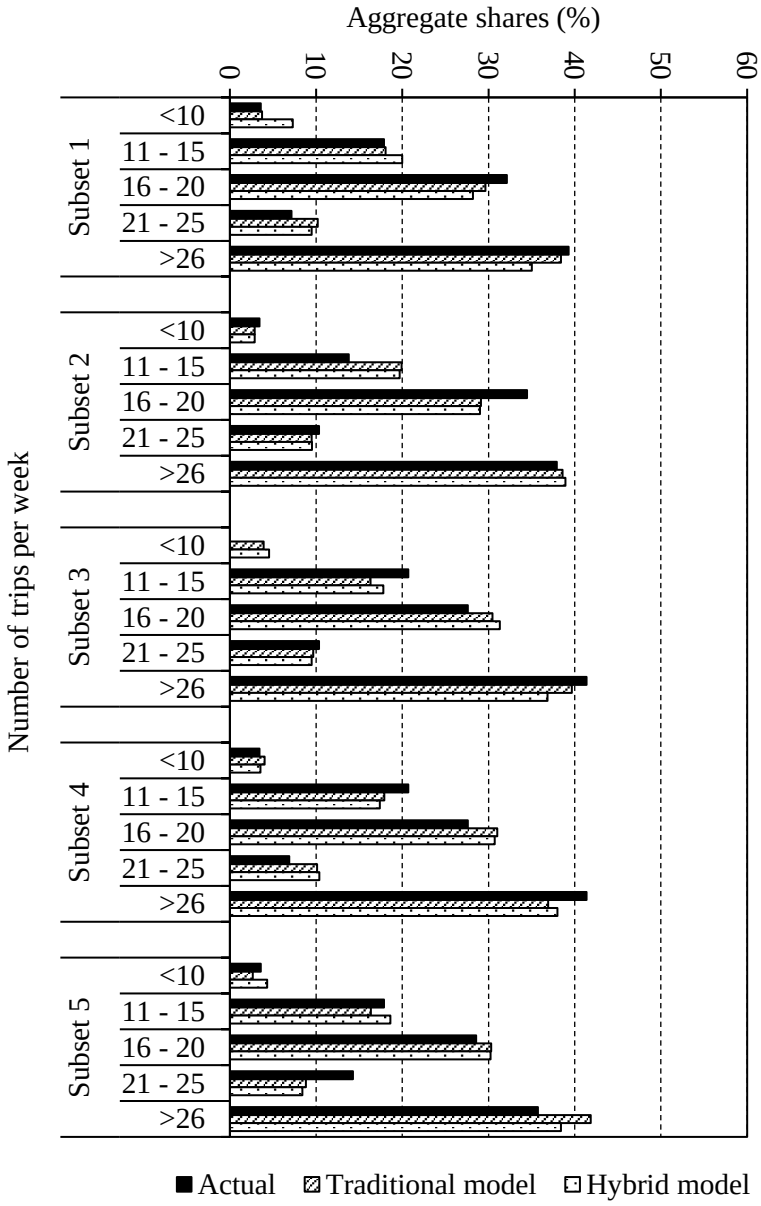


Figure 2-4 Trip generation model predictive performance in the validation subsets

As can be observed, both the actual and the predicted shares tend to follow a similar trend for both models albeit with observable differences across all the five subsets. The difference between the actual and predicted shares for both models is probably due to the use of weak explanatory variables. As mentioned, previous trip generation studies have shown that income and car ownership are some of the most important explanatory variables (e.g. Bwambale et al., 2015) and yet these were not considered in this study.

The predictive measures of fit for both models were computed and are presented in Table 2-5. At the aggregate level, the hybrid model performs better than the traditional model in three out of the five subsets in terms of the root mean square error. At the disaggregate level, the hybrid model performs better than the traditional model in four out of the five subsets in terms of the average probability of correct prediction, and the predictive rho-square. As mentioned earlier, the relatively better performance of the hybrid model could be in part due to allowing for heterogeneity through the probabilistic component. Nevertheless, these results prove that the proposed hybrid model is a feasible alternative to the traditional model, particularly where other reliable data sources are absent, thereby supporting the use of predicted demographics.

Table 2-5 Trip generation model measures of fit in the validation subsets

Validation subset	Root Mean Square Error		Average probability of correct prediction		Predictive rho-square	
	Traditional model	Hybrid model	Traditional model	Hybrid model	Traditional model	Hybrid model
Subset 1	1.81	3.41	0.273	0.287	0.102	0.121
Subset 2	3.69	3.65	0.287	0.301	0.122	0.132
Subset 3	3.04	3.59	0.302	0.304	0.190	0.202
Subset 4	3.16	2.97	0.303	0.279	0.171	0.129
Subset 5	3.84	3.03	0.280	0.282	0.107	0.120

2.8 Summary and conclusions

The paper demonstrates the feasibility of the hybrid framework to mitigate the challenges associated with the estimation and the application of trip generation models using mobile phone data. An examination of the parameter signs and the t-statistics for the differences between the parameters of the hybrid trip generation model (with predicted demographics) and a traditional model (with observed demographics) shows that both models capture the same trip generation behaviour, an indication that both models would lead to similar policy conclusions.

We also assess the performance of the traditional and the hybrid trip generation models using several measures of fit in five estimation and validation samples. For the estimation samples, we compare the final log-likelihoods while for the validation samples, we compare the root mean square error values (the predicted and actual shares), the predictive rho-square values, and the average probabilities of correct prediction. We find that the traditional model performs slightly better than the hybrid model during estimation and attribute this to the error-free measures of the socio-demographic variables in the traditional model with observed demographics. However, we find that the hybrid model generally performs better than the traditional model during validation in terms of the root mean square error values,

the predictive rho-square values, and the average probabilities of correct prediction. We attribute this improved performance to the possibility that the hybrid model allows for heterogeneity through the probabilistic component.

For demographic prediction, we find that the performance of the model is satisfactory. However, this being a secondary data set, there are limitations in the sample size and distribution that are beyond our control. For example, we note that some demographic groups have very small sub-sample sizes which could have affected the overall model performance. We therefore recommend further research into different ways of improving the demographic prediction component of the hybrid model by dedicated data collection efforts.

In practice, the proposed hybrid framework could be used where one has the demographic information, call detail records, and GSM mobility data for just a small representative section of willing users for the purposes of model calibration and anonymous CDR data for the full population. We note that GSM mobility data is generally discarded by mobile phone operators due to storage space constraints, however, it is possible to store such data for a small sub-sample of willing users. Once calibrated, the model only needs the phone usage characteristics of the individuals to be implemented and these can be derived from the anonymous CDRs of the entire population. The model can thus be applied for planning purposes, particularly where other reliable data sources are absent.

We conclude that the validation results serve as a proof-of-concept that having the demographics of a sub-sample of willing mobile phone users can make mobile phone data feasible for econometric travel behaviour modelling and travel demand estimation. Further, the proposed hybrid framework has promise in improving the modelling of the other stages of the 4 step model (e.g. mode choice, route choice, etc.) using mobile phone data by enriching them with probabilistic latent socio-demographic classes in the absence of observed ones.

Acknowledgements

We would like to thank the Economic and Social Research Council (ESRC) of the UK and Institute for Transport Studies, University of Leeds for funding this research. The authors also acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS. The research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia.

References

Aarthi, S., Bharanidharan, S., Saravanan, M. & Anand, V. Predicting customer demographics in a mobile social network. Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, 2011. IEEE, 553-554.

Agard, B., Morency, C. & Trépanier, M. 2006. Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes,* 39**,** 399-404.

Akin, D. & Sisiopiku, V. P. Estimating origin–destination matrices using location information from cell phones. Proc. 49th Annual North American Meetings of The Regional Science Association Int, Puerto Rico, 2002.

Ben-Akiva, M. E. & Lerman, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

Blumenstock, J. E. 2015. Calling for better measurement: Estimating an individual's wealth and well-being from mobile phone transaction records.

Blumenstock, J. E., Gillick, D. & Eagle, N. 2010. Who's calling? Demographics of mobile phone use in Rwanda. *Transportation*, 32**,** 2-5.

Bolla, R., Davoli, F. & Giordano, F. Estimating road traffic parameters from mobile communications. Proceedings 7th World Congress on ITS, Turin, Italy, 2000.

Borooah, V. K. 2001. *Logit and probit*, London, Sage Publications.

Brdar, S., Culibrk, D. & Crnojevic, V. Demographic attributes prediction on the real-world mobile data. Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK, 2012.

Bwambale, A., Choudhury, C. F. & Sanko, N. Modelling Car Trip Generation in the Developing World: The Tale of Two Cities. Transportation Research Board 94th Annual Meeting, 2015.

Chakirov, A. & Erath, A. 2012. Activity identification and primary location modelling based on smart card payment data for public transport.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

De Groote, A. 2005. GSM Positioning Control. *University of Fribourg, Switzerland***,** 13.

Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L. & Wang, D. 2016. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences***,** 201525443.

Dong, Y., Yang, Y., Tang, J., Yang, Y. & Chawla, N. V. Inferring user demographics and social strategies in mobile social networks. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014. ACM, 15-24.

Frias-Martinez, V., Frias-Martinez, E. & Oliver, N. A gender-centric analysis of calling behavior in a developing economy. AAAI Symposium on Aritificial Intelligence and Development, 2010.

Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. 2008. Understanding individual human mobility patterns. *Nature*, 453**,** 779-782.

Greene, W. H. & Hensher, D. A. 2010. *Modeling ordered choices: A primer*, Cambridge University Press.

GSMA Intelligence. 2015. *The Mobile Economy 2015* [Online]. Available: http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf [Accessed 26 July 2016].

Hasan, S., Zhan, X. & Ukkusuri, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, 6.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41**,** 260-271.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies,* 40**,** 63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A. & Willinger, W. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012. Acm, 239-252.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. & Laurila, J. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J. & Miettinen, M. The mobile data challenge: Big data for mobile computing research. Pervasive Computing, 2012.

Marschak, J. 1960. Binary Choice Constraints on Random Utility Indications. *In:* ARROW, K. (ed.) *Stanford Symposium on Mathematical Methods in the Social Science*. Stanford, California: Stanford University Press.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)***,** 109-142.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics***,** 105-142.

McKelvey, R. D. & Zavoina, W. 1975. A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology,* 4**,** 103-120.

Mo, K., Tan, B., Zhong, E. & Yang, Q. Report of task 3: your phone understands you. Nokia mobile data challenge 2012 workshop, Newcastle, UK, 2012. Citeseer, 18-19.

Ortúzar, J. D. D. & Willumsen, L. G. 2011. *Modelling transport*, John Wiley & Sons.

Pan, C., Lu, J., Di, S. & Ran, B. 2006. Cellular-based data-extracting method for trip distribution. *Transportation Research Record: Journal of the Transportation Research Board***,** 33-39.

Song, C., Koren, T., Wang, P. & Barabási, A.-L. 2010. Modelling the scaling properties of human mobility. *Nature Physics,* 6**,** 818-823.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A. & González, M. C. 2015. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*.

Vickerman, R. & Barmby, T. 1985. Household trip generation choice—Alternative empirical approaches. *Transportation Research Part B: Methodological,* 19**,** 471-479.

White, J. & Wells, I. Extracting Origin Destination Information from Mobile Phone Data. Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), March 2002 London. IET, pp. 30 - 34.

Wu, L., Zhi, Y., Sui, Z. & Liu, Y. 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one,* 9**,** e97010.

Ying, J. J.-C., Chang, Y.-J., Huang, C.-M. & Tseng, V. S. 2012. Demographic prediction based on users mobile behaviors. *Mobile Data Challenge*.

Zheng, V. W., Zheng, Y., Xie, X. & Yang, Q. Collaborative location and activity recommendations with gps history data. Proceedings of the 19th international conference on World wide web, 2010. ACM, 1029-1038.

Chapter 3
Modelling long-distance route choice using mobile phone call detail record data: a case study of Senegal

Andrew Bwambale[*], Charisma F. Choudhury[*], Stephane Hess[*]

Abstract

Over the last two decades, Global Positioning System (GPS) data has been widely used for route choice modelling. However, such studies are often expensive, thereby leading to small sample sizes and increased risks of sampling bias. On the other hand, call detail records (CDRs) are more readily available for millions of users due to the growing mobile phone penetration rates worldwide and could serve as a low-cost alternative. This motivates this research where we investigate the potential of using CDR data for route choice modelling. We analyse the limitations of CDR data and propose techniques for inferring the chosen routes or subsets of the likely routes from partial CDR trajectories. Considering the anonymous nature of CDR data, we address issues of choice set determination and data fusion prior to model estimation. Due to the partial nature of CDR trajectories, route choice is observed at disaggregate and aggregate levels, which prompts us to adapt the broad choice framework to route choice modelling. Intuitive model results are obtained and used to estimate the value of travel time, which is found to be realistic for Senegal. The research findings are useful for developing countries where budgetary constraints on transport studies are common.

Keywords: Route choice behaviour, Broad choice, Mobile phone data, Call detail records, Value of travel time

[*] Choice Modelling Centre, Institute for Transport Studies, University of Leeds (UK)

3.1 Introduction

The modelling of route choice behaviour for long journeys, inter-city and inter-regional trips has been an important aspect of transport research for several decades, however, the number of studies in this field using revealed preference (RP) data is still low (e.g. Hess et al., 2015, Ben-Akiva et al., 1984). This is partly due to the demanding data collection requirements for such studies.

Traditional RP data collection approaches in this context rely on interviews and paper or web-based questionnaire surveys where individuals are asked to describe the routes taken for particular trips (e.g. Vrtic et al., 2006, Ramming, 2002). These data collection techniques are generally expensive, which leads to limited sample sizes, thus increasing the risk of sampling biases. This problem is particularly prevalent in developing countries where stringent budget constraints on transport studies are common. Moreover, traditional data collection is often affected by low response rates and reporting errors which can lead to biased model estimates (e.g. Groves, 2006).

The last few decades have seen the emergence of various technologies that enable the passive collection of mobility trajectories, mitigating the burden of route choice data collection. This has led to numerous route choice studies based on Global Positioning System (GPS) data, primarily from navigational devices (Li et al., 2018, Hess et al., 2015, Broach et al., 2012, Bierlaire and Frejinger, 2008), and more recently from smartphones (Bierlaire et al., 2010, Papinski et al., 2009). Although the incorporation of Assisted-GPS (A-GPS)[3] features in most smartphones has significantly improved the accuracy of GPS locations, data generation strongly relies on smartphone ownership, internet connectivity, and data storage capacity, which leads to small sample sizes as seen in most related studies (e.g. Nitsche et al., 2014, Bierlaire et al., 2013, Nitsche et al., 2012, Bierlaire et al., 2010) thereby increasing the risk of sampling bias.

This problem can be overcome by taking advantage of the large-scale anonymous datasets that are already being passively collected by operators for different purposes, and applying these to transport studies. Such datasets have already yielded promising results in various mobility studies. Examples include; social media data (Hawelka et al., 2014, Hasan et al., 2013), smart card data (Chakirov and Erath, 2012, Agard et al., 2006), and network-generated mobile phone data such as Call Detail Records (CDRs)[4] and Global System for Mobile communications (GSM)[5] data (Çolak et al., 2015, Jiang et al., 2013, Schlaich et al., 2010). However among these, network-generated mobile phone data is particularly a promising source due to the high mobile phone penetration rate worldwide (GSM Association, 2017).

A review of the literature shows that there have been a few route identification studies using network-generated mobile phone data (e.g. Nie et al., 2015, Leontiadis et al., 2014, Hoteit et al., 2014, Schlaich et al., 2010), however, most of these studies end on route identification

[3] A-GPS data comprises of triangulated mobile phone positions obtained by enhancing standalone GPS data using neighbouring cell tower locations to obtain more accurate and precise positions in poor satellite signal conditions.

[4] CDR data reports the time stamped locations of communication events (i.e. voice calls, text messages, and data calls) as well as the details of the request (i.e. the duration and direction).

[5] GSM data reports the IDs of all the GSM cells traversed by an active mobile phone at regular time intervals (irrespective of the calling or texting patterns of the users).

and do not attempt to investigate the factors affecting route choice behaviour. At the moment, only Schlaich (2010) combines GSM trajectories with traffic state information to analyse the influence of variable message signs (VMS) and other factors on route choice. However, the success of Schlaich's study could in part be attributed to the use of GSM data, which is semi-continuous in nature as opposed to CDR data which is discontinuous. Although the discontinuous nature of CDR locations presents serious trajectory identification challenges, the data is more readily available for millions of users at zero/minimal costs as it is stored by operators for billing purposes. Therefore, route choice models based on CDR data can be practically useful in both developed and developing countries. This motivates this research where we focus on using CDR data for modelling route choice behaviour.

However, it is important to underscore the practical challenges that stem from the use of CDR data, and how this impacts our work. Given the discontinuous nature of CDR data, we are only able to observe the partial trajectory of a user depending on their phone usage rate during a particular trip. For very close O-D pairs, it is likely that a user may travel from origin to destination without using their phone, thus making it impossible to even capture the partial trajectories. However, for distant O-D pairs, there is an increased possibility that a user will use his/her phone at different points during the trip thus enabling the capture of his/her partial trajectory. For this reason, our study focuses on long-distance inter-regional route choice. Another reason for considering long-distance route choice is that there are usually fewer and widely spaced alternative routes, which can easily be identified despite the low location accuracy of CDR data.

Furthermore, the locations at which the phones are used also matters since some location areas can be associated with more than one possible route. In such cases, it is not possible to precisely infer the chosen routes from the partial trajectories, rather, route choice is observed at a broad sub-group level (e.g. northern, southern etc.), where each sub-group comprises of a small set of possible routes. This prompts us to adapt the broad choice modelling framework, developed in the context of vehicle type choice (Wong, 2015) to route choice modelling using noisy CDR data. We note that this may be problematic in dense inter-urban networks, where it would be difficult to identify a small enough subset of possible routes using a few CDR locations. However, with the increasing trend of mobile internet usage (Gerpott and Thomas, 2014), the frequency of CDR locations is likely to improve significantly in the near future, which adds further to the timeliness of the present paper.

This paper addresses issues of route identification, choice set determination and data fusion (with conventional and non-conventional data sources) prior to route choice model development. Although CDR data is only able to capture the partial trajectories of frequent phone users, the samples are usually large, thus increasing the possibility that they are representative enough to capture rational route choice behaviour. The need to investigate this assertion forms the basis of our validation exercise. The developed models are used to estimate the value of travel time (VTT) for Senegal (the study area), yielding reasonable estimates. The study is timely in the sense that it extends the application of CDR data beyond travel pattern visualisation to econometric modelling of travel behaviour. This could motivate reliable and low-cost policy formulation in different contexts. While we use Senegal as a case study, this research is beneficial to other developing countries with budgetary constraints on transport studies.

The rest of the paper is arranged as follows; section 3.2 presents a review of relevant literature, section 3.3 presents the data description, section 3.4 presents the data processing conducted, section 3.5 presents the modelling framework, section 3.6 discusses the model results, while section 3.7 presents the summary and conclusions.

3.2 Literature review

This section briefly reviews the literature on the applications of mobile phone data in transport studies, as well as different models of route choice.

3.2.1 Previous applications of mobile phone data to transportation studies

The last few decades have seen significant research effort in the application of large-scale mobile phone data to transportation studies. Such data has been widely applied in the development of human mobility models (e.g. Deville et al., 2016, Isaacman et al., 2012, Song et al., 2010), estimation of trip rates (e.g. Çolak et al., 2015), development of origin-destination matrices (e.g. Çolak et al., 2015, Iqbal et al., 2014, White and Wells, 2002), travel mode detection (e.g. Qu et al., 2015, Wang et al., 2010, Reddy et al., 2008), and traffic model calibration (e.g. Bolla et al., 2000).

However, we place focus on studies related to route identification. A few of these studies have used GSM data, which reports the complete mobile phone location area sequences of each user, thus enabling the easy identification of routes through sequence matching (e.g. Tettamanti et al., 2012, Schlaich et al., 2010) and probabilistic methods such locality-sensitive hashing and graph clustering (e.g. Görnerup, 2012). Instead, most studies have focussed on analysing the potential of CDR data, which is more widely available and yet challenging in this context. For example, Doyle et al. (2011) use the virtual cell paths technique to extract user trajectories from CDR data, and generate the kernel density paths for different routes to validate their findings. Saravanan et al. (2011) analyse the spatial and temporal information of CDR events over a long period of time to establish the daily routines and routes of the users. Hoteit et al. (2014) join subsequent triangulated CDR locations using linear, cubic, and nearest - neighbour interpolation to model the potential trajectories. Leontiadis et al. (2014) calculate the weights for each road segment within the cell areas linked to a user's communication events and determine the shortest weighted path for a given OD pair. Nie et al. (2015) mark each route with a subset of *k* optimal cell handover sequences extracted from the full set of possible handover sequences and matches these with those observed in the cell phone hand over data (similar to CDR data) based on the degree of similarity.

In this study, we follow an approach slightly similar to that in Nie et al. (2015), however, instead of using a similarity index, we pursue the idea of unique and shared location area sequences in the context of broad choice modelling as explained later.

3.2.2 Existing route choice models

The vast majority of route choice models belong to the family of discrete choice models (see Ben-Akiva and Lerman, 1985 for details), with the multinomial logit (MNL) model (McFadden, 1974) being the most widely used. However, the MNL model is affected by the irrelevance of independent alternatives (IIA) property, which can be problematic for highly overlapping routes (Ramming, 2002). This has motivated the development of more

advanced route choice models to address this challenge. Examples include the nested recursive logit model (Mai et al., 2015), the c-logit model (Cascetta et al., 1996), the path size logit model (Ben-Akiva and Ramming, 1998), the link nested (cross-nested) logit model (Vovsha and Bekhor, 1998), and the multinomial probit model with logit kernel (Daganzo et al., 1977). Details of how some of these models overcome the overlapping route problem are discussed in section 3.5.3 of this paper.

An important point to note is that the complexities of route choice modelling go beyond the overlapping route problem. Choice set generation is a key challenge, especially in highly overlapping dense urban networks, where several alternative routes can be possible, and yet individuals do not consider all the alternatives while making choices (Prato, 2009). Several choice set generation methods have been proposed in the literature including, the k-shortest path algorithms (e.g. Shier, 1979, Bellman and Kalaba, 1960), the labelling approach (Ben-Akiva et al., 1984), link elimination approaches (e.g. Azevedo et al., 1993, Bellman and Kalaba, 1960), link penalty approaches (e.g. Rouphail et al., 1995, De La Barra et al., 1993), simulation approaches (e.g. Sheffi and Powell, 1982), doubly stochastic generation functions (e.g. Nielsen, 2000), constrained enumeration methods (e.g. Prato and Bekhor, 2006), and probabilistic methods (e.g. Cascetta and Papola, 2001, Manski, 1977).

However, since the focus of this paper is long distance trips, where the alternatives are usually few in number, choice set determination is more straightforward as discussed later in section 3.4.2 of this paper.

3.3 Data

This study uses CDR data collected from Senegal as part of the Orange Data for Development (D4D) challenge (de Montjoye et al., 2014).

3.3.1 Study area

Senegal is located in West Africa with a population of approximately 13.5 million according to the 2013 population census (ANSD, 2016).

Road transport accounts for over 99% of all passenger travel (World Bank, 2004). The only long-distance train service (the Dakar-Niger line) was discontinued in May 2010 (Imedia and Calao Production, 2013).

The country has a sparse national road network (see Figure 3-1), and for some O-D pairs, there is only one feasible alternative, making them unsuitable for route choice modelling. This study ignores such O-D pairs and only considers those where alternative routes exist.

In total, twelve distant O-D pairs are considered, and these are; Dakar-Bakel, Dakar-Matam, Thies-Bakel, Thies-Matam, Diourbel-Bakel, Diourbel-Matam, and the corresponding O-D pairs for the reverse directions as shown in Figure 3-1.

The long travel times between these regions increase the possibility of capturing the users' partial trajectories as explained earlier.

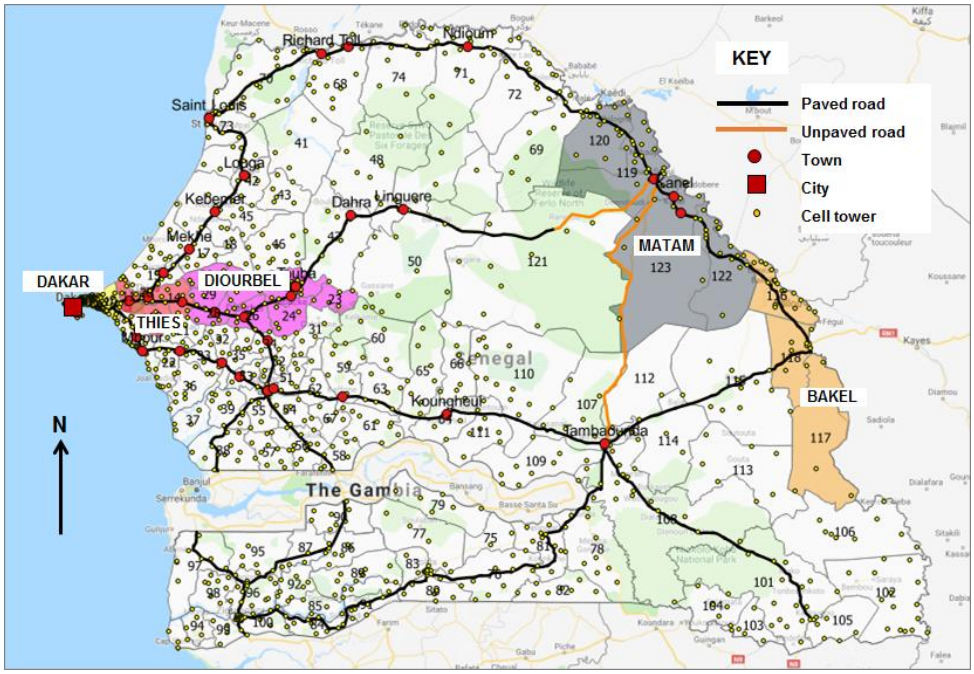


Figure 3-1 Study area (Google Maps, 2017b, Worldatlas, 2017, ArcGIS, 2013)

3.3.2 CDR data

The CDR data was collected between January and December 2013 and aggregated to the arrondissement (district) level by the data provider. The geographical location of the arrondissements is presented in Figure 3-1 where we also show the tower locations for illustration purposes.

The original CDR data comprised of 9 million unique users (67% of the study area population). This was pre-processed to retain frequent phone users (i.e. those with interactions on 75% of the days in a year) and randomly split into smaller monthly rolling sub-samples made available for research (see de Montjoye et al., 2014 for details). The user IDs in each sub-sample are anonymised to prevent possible re-identification across the different months.

The data for each month comprises of about 150,000 users. Taking the most commonly observed arrondissement for each user during the month as their home district, the monthly population sampling rate ranged from 2.4% in Dakar (the capital) to 0.4% in the rural regions. On average, these users together generated over 40 million records per month (see an excerpt of the CDR data in Table 3-1a). The data is reduced to remove duplicate records resulting in the processed arrondissement visitation data presented in Table 3-1b.

The overall level of user mobility is illustrated in Figure 3-2. As shown, most users visited less than three unique arrondissements per month. The low levels of inter-arrondissement mobility led to the capture of few trajectories as reflected in the final sample size (see Section 3.4.2).

Table 3-1a Excerpt of the raw CDR data

Anonymised User ID	Timestamp	Arrondissement ID[6]
130599	13-01-02 20:10	25
130599	13-01-13 13:10	7
130599	13-01-19 23:50	19
130599	13-01-19 23:50	19
130599	13-01-22 01:30	2
130599	13-01-22 01:30	2
130599	13-01-28 20:20	4
130599	13-01-28 20:20	4
130599	13-01-29 19:40	4
130599	13-01-29 19:50	4
130599	13-01-29 20:00	4
130599	13-01-29 20:40	4
130599	13-01-29 21:20	4
130599	13-01-29 21:50	4
130599	13-01-29 21:50	4
130599	13-01-29 21:50	4

Discarded from the data

Table 3-1b Excerpt of the processed arrondissement visitation data

Anonymised User ID with a monthly identifier (e.g. January)	Arrondissement ID	1st observation	Last observation
130599.01	25	13-01-02 20:10	13-01-02 20:10
130599.01	7	13-01-13 13:10	13-01-13 13:10
130599.01	19	13-01-19 23:50	13-01-19 23:50
130599.01	2	13-01-22 01:30	13-01-22 01:30
130599.01	4	13-01-28 20:20	13-01-29 21:50

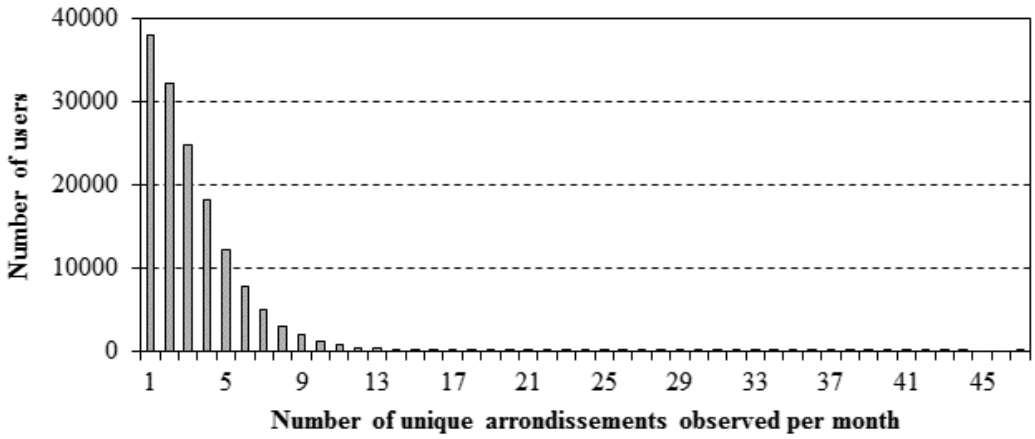


Figure 3-2 Average monthly arrondissement observation frequency distribution

[6] The geographical locations of the arrondissement IDs are presented in Figure 3-1

Although it is difficult to detect false tower jump movements in aggregate CDR data (Çolak et al., 2015, Iqbal et al., 2014), this is not a big factor for distant O-D pairs where the origin and destination arrondissements have been grouped into regions.

3.4 Data preparation for analysis

This section describes the analysis carried out on the processed arrondissement visitation data in Table 3-1b to identify the routes followed, as well as the processes of estimating the route attributes.

3.4.1 Route identification

The route identification process is summarised in Figure 3-3. This is divided into two main stages as described in the subsequent sections.

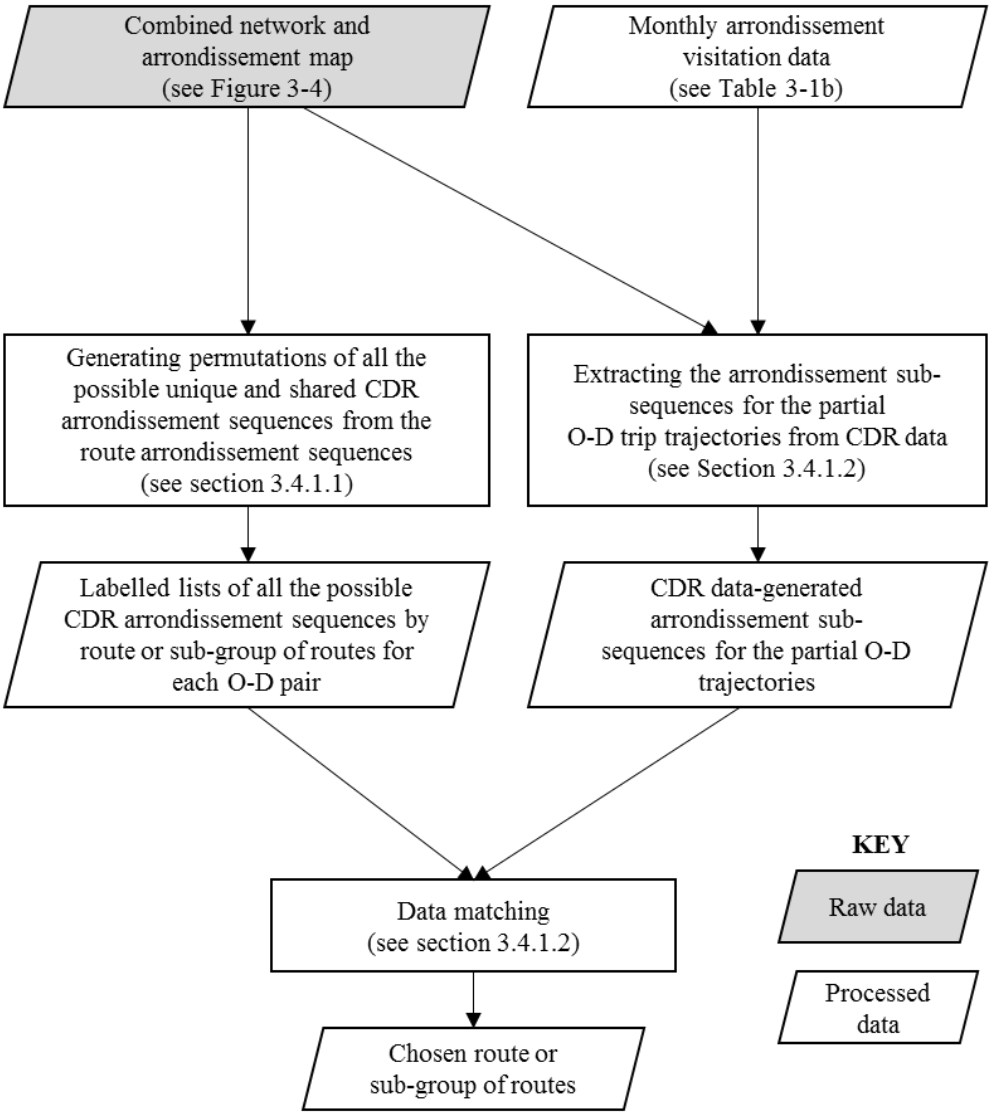


Figure 3-3 Summary of the route identification process

3.4.1.1 Generation of unique and shared CDR arrondissement sequences

Route arrondissement sequences (which are extracted from maps) show the order of all the arrondissements traversed by a particular route between a given O-D pair. On the other hand, CDR arrondissement sequences (which are extracted from the CDR data) show the order of the arrondissements in which a user used his/her phone during the trip, and are subsets of the route arrondissement sequences.

For any given trip along a particular route, several possible CDR arrondissement sequences can be observed depending on the number and the location of the CDR events. These can be obtained by generating permutations of different sizes based on the route arrondissement sequence (in which order matters and no repetitions are allowed). However, since most of the O-D pairs have overlapping routes, some of the CDR arrondissement sequences can be linked to more than one route if all the intermediate CDR events occurred along the shared sections. In this case, it would only be possible to observe the subset of routes that were potentially followed (see illustration in Figure 3-4 where the blue areas indicate the shared arrondissements, while the grey and green areas indicate the unique arrondissements for the Northern 1 and 2 routes respectively).

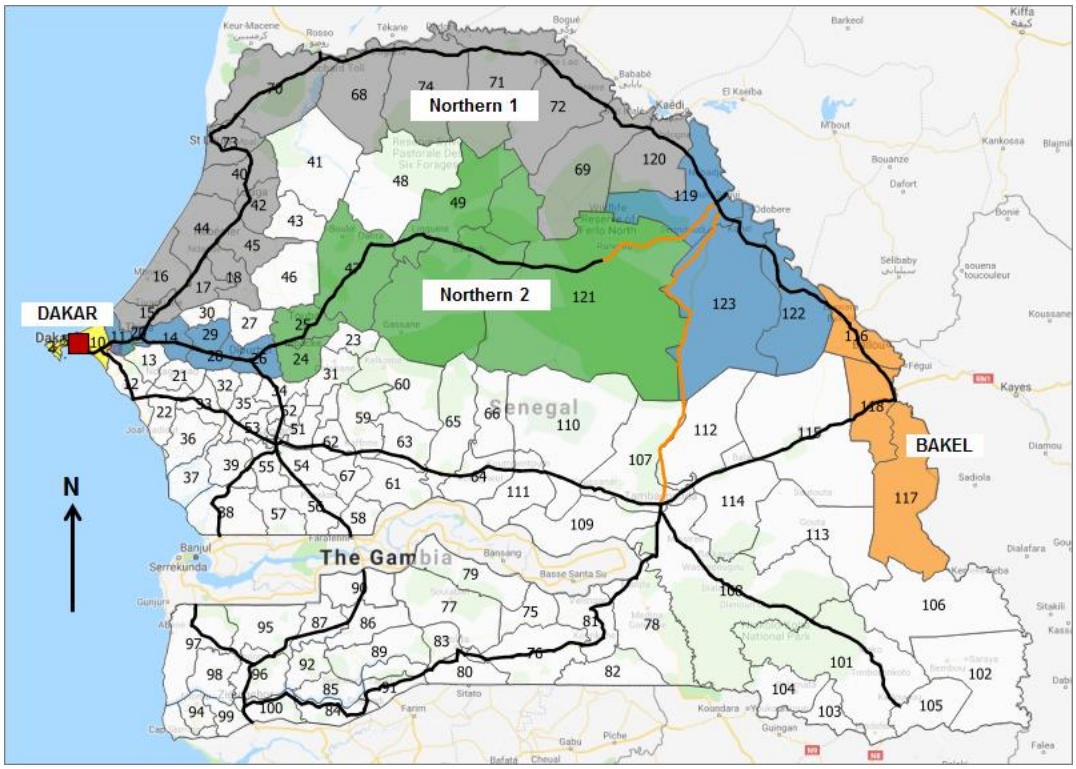


Figure 3-4 Arrondissement paths (Dakar-Bakel O-D pair as an example)

The generated CDR arrondissement sequences linked to each route were cross-referenced to identify the permutations linked to unique routes (i.e. unique CDR arrondissement sequences) and those shared across multiple routes (i.e. shared CDR arrondissement sequences). The outcome of this analysis was a list of all the possible CDR arrondissement sequences labelled with the associated routes or sub-groups of the possible routes (see example in Table 3-2).

3.4.1.2 Extraction the O-D pair trip trajectories from CDR data

The processed arrondissement visitation data for each user (see excerpt in Table 3-1b) was analysed to extract sub-sequences linked with possible trips between the regions of interest following the criteria below;

- The first and the last arrondissements in the sub-sequence must be located within different regions of interest, and the user must not be observed in an upstream or a downstream region of interest within the same day for origins and destinations respectively;
- The dwell time in the origin and the destination regions of interest must be longer than that required to directly traverse each of them to increase the possibility that these are the trip start and end locations. A user needs to use his/her phone at least twice in each of these regions to calculate the dwell time;
- The intermediate arrondissements in the sub-sequence must all be associated with one of the defined arrondissement/corridors paths (see Figure 3-4) to ensure only direct trips are retained; and
- The timestamp difference between the origin and the destination must not exceed 24 hours, which is used as an upper limit to distinguish between users with direct trips but delay to use their phones on arrival, and those with intermediate destinations, thereby arriving late.

The extracted sub-sequences meeting all the above criteria were either assigned to unique routes or sub-groups of the possible routes by cross-referencing with the labelled lists generated in Section 3.4.1.1. Table 3-2 presents an excerpt of the route assignment data.

Table 3-2 Excerpt of the route assignment data

Anonymised User ID with a monthly identifier (e.g. January)	CDR Trajectory	Route/ Broad sub-group
131891.01	Dakar-11-C1-Bakel	Northern 1
131891.01	Dakar-C1-Bakel	Northern 1
132801.01	Dakar-122-Bakel	Northern (Northern 1/ Northern 2)*
132801.01	Dakar-28-C2-123-Bakel	Northern 2
132801.01	Dakar-C1-123-Bakel	Northern 1

* CDR trajectory can belong to both the Northern 1 and Northern 2 routes

In this data, 70% comprised of unique assignments, while 30% comprised of broad assignments. Since this is a scenario where for some users and/or trips, we know the chosen route at a more disaggregate level than for others, we use the Broad Choice Modelling Framework (Wong, 2015) to analyse route choice behaviour.

3.4.2 Estimation of route attributes

For model estimation, it is critical to determine the choice set and the attributes of the alternatives. We assumed that the choice set is comprised of the routes that have ever been chosen by the different users. Routes not chosen by any user for the whole year were excluded from the choice set. On average, each origin-destination pair had four alternatives. Given that the total dataset is comprised of 9,453 records from 6,497 users, this is not a very restrictive assumption. Given the choice sets, we reviewed previous studies to identify the

attributes typically used in route choice models and their availability status for Senegal as summarised in Table 3-3. Although data on six explanatory variables was available, the final model specification contains three explanatory variables only as the inclusion of the other variables led to correlation problems and/or illogical model results. A detailed explanation of the variable specification process is presented in Section 3.6.1. The subsequent sections summarise the processes of estimating some of these attributes.

Table 3-3 Attributes typically used in route choice models

Attribute	Sample references	Data availability	Remarks
Individual socio-demographics	(Ramming, 2002, Zhang and Levinson, 2008)	×	Mobile phone data is usually anonymous
Travel time	(Hess et al., 2015, Ramming, 2002, Ben-Elia and Shiftan, 2010)	✓	Can be derived from traditional data or maps
Travel cost	(Hess et al., 2015)	✓	Can be estimated using the vehicle operating costs
Distance	(Hamerslag, 1981, Bitzios and Ferreira, 1993)	✓	Can be calculated from maps
Scenic characteristics	(Zhang and Levinson, 2008, Ben-Akiva et al., 1984)	✓	Can be derived from maps
Safety (e.g. presence of black spots)	(Ben-Elia and Shiftan, 2010, Ben-Akiva et al., 1984)	×	Data could not be obtained
Urban developments along the route	(Zhang and Levinson, 2008, Ben-Akiva et al., 1984)	✓	Data can be obtained from maps
Time or distance on uninterrupted flow facilities (e.g. freeways)	(Bierlaire and Frejinger, 2008, Ramming, 2002)	×	No such facilities between the regions of interest at the time of data collection
Traffic congestion	(Bitzios and Ferreira, 1993)	×	Data could not be obtained
Road quality (e.g. road surface conditions)	(Ben-Akiva et al., 1984)	✓	Data from the national roads agency is available
Road signs (e.g. direction signs)	(Wootton et al., 1981)	×	Data could not be obtained

3.4.2.1 Link length and surface attributes

Data on the link lengths and surface attributes (i.e. paved or unpaved) was derived from the Senegal roads GIS layer (ArcGIS, 2013). This was updated to reflect the situation in 2013 relying on road condition reports sourced from government and other relevant websites (Ageroute Senegal, 2017, ANSD, 2017, Logistics Cluster, 2013).

3.4.2.2 Travel time

Travel time cannot be reliably estimated from the CDR data as users do not necessarily use their phones at the moment of departure or arrival. The typical travel times for most links in 2013 were obtained from the website of Logistics Capacity Assessment (Logistics Cluster, 2013). For links not covered by this website, we relied on Google Maps (Google Maps, 2017a).

It may be noted that we used the same average travel time for all the users along a particular route between a given O-D pair. A better approach would have been to estimate user-specific travel times based on the corresponding actual departure or arrival times.

Although such information can be obtained from Google Maps using the directions tool (Google Maps, 2017a), it is difficult to observe the actual departure or arrival times from CDR data due to its discontinuous nature as mentioned earlier, moreover, Google Maps does not report the nation-wide travel time variability for Senegal at the moment.

3.4.2.3 Travel cost

Travel cost was estimated in terms of the vehicle operating costs (VOCs) per user (i.e. fuel and non-fuel costs).

After a review of several VOC estimation techniques, we settled for the HDM-III model (Watanatada et al., 1987) due to its applicability to developing countries and input data availability. The HDM-III model is an earlier version of the more advanced HDM-4 (Kerali, 2000), which we could not use due to input data constraints.

The HDM-III model relies on vehicle calibration data (where we used default values) and other basic input data (see Table 3-4) to estimate the VOCs for each vehicle type. The model works by defining link-specific relationships between the International Roughness Index - IRI (Sayers et al., 1986) and speed, and using the IRI values at the respective average link speeds (derived from Sections 3.4.2.1 and 3.4.2.2) to estimate the link-specific VOCs, which are summed to estimate the route VOCs.

The estimated route VOCs need to be converted to person costs. Given the anonymous nature of CDR data, we use information on the typical occupancy rates and mode shares (see Table 3-5) to estimate the weighted average VOCs per user for each route.

As was the case with travel time, we used the same average travel cost for all the users along a particular route between a given O-D pair. An improved approach would have been to estimate user-specific travel costs based on the corresponding travel speeds, however, this was not possible due to difficulties in obtaining the user-specific travel times as discussed in the previous section.

Table 3-4 HDM-III basic input data

Input	Measure	Unit	Car	Inter-urban taxi	Minibus	Bus	Source
Terrain type	Rise &fall	m/km	Estimated directly for each link using Google Earth and Auto CAD Civil 3D				(Autodesk, 2017, Google Earth 7.1.8.3036, 2016)
	Horizontal curvature	deg/km					
Desired max speed	Desired max speed	km/hr	90	90	90	90	(WHO, 2016)
Economic unit costs (Excluding taxes)*	Vehicle cost price	$	19,452	27,098	49,474	67,465	(ADF, 2011)
	Fuel type	NA	Petrol	Diesel	Diesel	Diesel	(ADF, 2011)
	Fuel costs	$/litre	1.017	0.727	0.727	0.727	(ADF, 2011)
	Lubricants	$/litre	4.688	6.466	6.466	6.466	(ADF, 2011)
	Tyres + tubes	$	83.36	83.36	83.36	187.53	(ADF, 2011)
	Maintenance costs	$/hour	0.284	0.310	0.310	0.541	(ADF, 2011)
	Crew costs	$/hour	0	0.511	0.511	0.511	(ADF, 2011)
	Interest rate	%	5.4	5.4	5.4	5.4	(World Bank, 2017b)
Utilisation	Mileage per year	km	25,000	50,000	50,000	60,000	(ADF, 2011)
	Hour driven per year	hours	350	750	750	1250	(ADF, 2011)
Service life	Service life	years	12	12	12	12	(ADF, 2011)
Gross vehicle weight	Gross vehicle weight	tons	1.2	2.0	3.0	11	(Watanatada et al., 1987)

* Prices adjusted for 2010-2013 inflation and the 2013 USD exchange rate (World Bank, 2017c, World Bank, 2017a)

Table 3-5 Typical occupancy rates and mode shares in Senegal (World Bank, 2004)

Vehicle type	Average number of passengers	Passengers per km	Mode share (%)
Cars	3	1746.7	0.183
Interurban Taxi	7	1830.4	0.191
Minibus	14	2149.6	0.225
Buses	25	3838.6	0.401

3.4.2.4 Other attributes

The other attributes considered were the scenic characteristics and the urban developments along each route. Scenic variables were estimated in terms of route lengths traversed through nature reserves (Google Maps, 2017b), while urban developments were reflected as the number of towns along each route as shown in Figure 3-1 (Worldatlas, 2017).

3.5 Modelling framework

We use discrete choice models in this study since route choices are discrete and mutually exclusive. To develop these models, we apply the random utility theory (Marschak, 1960), a well-established approach for estimating discrete choice models.

3.5.1 Basic model

Suppose U_{nr} is the utility of choosing route r by individual n. This can be expressed as;

$$U_{nr} = V_{nr} + \varepsilon_{nr} \quad (3\text{-}1)$$

Where V_{nr} and ε_{nr} are the systematic and the random parts utility respectively. The systematic utility is a function of the observed route attributes, and may be exprssed as $V_{nr} = \beta' X_{nr}$, where X_{nr} is a vector of the attributes of route r for individual n and β is a vector the model parameters. We assume that the random term ε_{nr} is independently and identically distributed across the alternatives following a type I extreme value distribution, and use the Multinomial Logit (MNL) model to estimate the route choice probabilities as follows (see McFadden, 1974 for details);

$$P_n(r) = \frac{\exp(V_{nr})}{\sum_{r^* \epsilon C_n} \exp(V_{nr^*})} \quad (3\text{-}2)$$

Where, $P_n(r)$ is the probability of individual n choosing route r, and C_n is the choice set. Given the route choice probabilities, the model parameters can be estimated by maximising the log-likelihood function below;

$$LL = \sum_n \sum_r \left[Z_{nr} * \ln\left(P_n(r)\right) \right] \quad (3\text{-}3)$$

Where Z_{nr} is a dummy variable, which is equal to 1 if and only if user n chooses route r.

3.5.2 Accounting for broad choices

The log-likelihood function in Equation 3-3 assumes that all the route choices are uniquely observed, and is inadequate for the current scenario, where we also have broad sub-group choices. Therefore, we use the broad choice modelling framework proposed by Wong (2015) to account for this situation.

In the broad choice framework, the choice probabilities of the broad sub-groups are expressed as a sum of the choice probabilities of the member alternatives. For example, the choice probability of the 'Northern' broad sub-group is the sum of the 'Northern 1' and the 'Northern 2' route choice probabilities (see Figure 3-4 and Table 3-2). The joint

probabilities of the broad sub-groups capture the aggregate shares at the unique route choice level using the relative probabilities of the constituent routes.

The goal of model estimation is to maximise the probabilities of both the observed routes and the broad sub-groups for users with unique and broad choices respectively. The log-likelihood function is specified as follows (Wong, 2015);

$$LL = \sum_{n} \sum_{b} \left[Z_{nb} * \ln \left(\sum_{r \epsilon S_b} P_n(r) \right) \right] \quad (3\text{-}4)$$

Where S_b is a set comprising of the routes in broad category b. For uniquely assigned trips, set S_b comprises of only one alternative. Z_{nb} is a dummy variable, which is equal to 1 if and only if user n is associated with category b.

3.5.3 Accounting for overlap

A major weakness of the MNL model (Equation 3-2) is the IIA property, which could lead to illogical route choice probabilities for highly overlapping routes as is the case in this study. This is illustrated using the overlapping route problem (Ramming, 2002, Cascetta et al., 1996) in Figure 3-5.

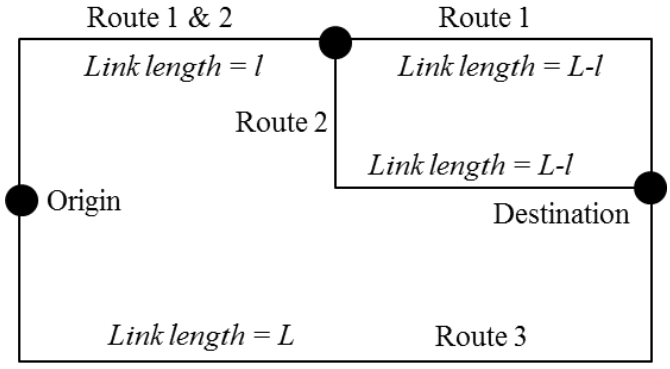


Figure 3-5 Overlapping route problem (Cascetta et al., 1996, Ramming, 2002)

Here, all three routes have the same total length L, however, routes 1 and 2 follow the same alignment for length l followed by distinct sections, each of length $L - l$. The MNL model predicts equal shares for each of the routes irrespective of the overlap length. However, as the overlap increases (as l tends to L), it becomes difficult to distinguish between routes 1 and 2, and it is expected that route 3 will take a share of 50%, while routes 1 and 2 will each take a share of 25%.

Various models accounting for overlap were presented in Section 3.2.2,, however, this study only uses the c-logit (Ramming, 2002, Cascetta et al., 1996) and the path size logit (Ben-Akiva and Ramming, 1998) models for illustration purposes as the modelling scenario did not require complex formulations. In general, these models are modifications of the MNL model, where the systematic route utilities are adjusted using certain correction factors as follows;

$$P_n(r) = \frac{\exp(V_{nr} + \tau_{nr})}{\sum_{r^* \epsilon C_n} \exp(V_{nr^*} + \tau_{nr^*})} \quad (3\text{-}5)$$

Where τ_{nr} is the systematic utility correction factor for route r.

3.5.3.1 C-logit model

For the c-logit model, the correction factor τ_{nr} is a commonality factor (Cascetta et al., 1996). Different possible specifications have been proposed, however, this study uses the following common specification (Ramming, 2002, Cascetta et al., 1996);

$$\tau_{nr} = \beta_{CF} \ln \left[\sum_{r^* \epsilon C_n} \left(\frac{L_{rr^*}}{\sqrt{L_r L_{r^*}}} \right)^{\gamma_{CF}} \right] \quad (3\text{-}6)$$

Where $L_{rr*}$is the overlap length between routes r and r^*, L_r and L_{r*} are the total lengths of routes r and r^* respectively, β_{CF} and γ_{CF} are the unknown parameters to be estimated. From Equation 3-6, the ratio in the brackets is proportional to the degree of overlap, while the corresponding logarithm has an inverse negative relationship. Thus β_{CF} and γ_{CF} are expected to have negative and positive signs respectively to allow for the positive adjustment of route utility with decreasing overlap (Ramming, 2002).

3.5.3.2 Path size logit model

For the path size logit model, the correction factor τ_{nr} is a path size term, which is computed as the weighted average of the constituent link sizes. The specification adopted for this study is as follows (Ben-Akiva and Ramming, 1998);

$$\tau_{nr} = \beta_{ps} \ln \left[\sum_{a \epsilon \Gamma_r} \left(\frac{l_a}{L_r} * \frac{1}{N_{ar}} \right) \right] \quad (3\text{-}7)$$

Where $1/N_{ar}$ is the inverse of the number of routes sharing the link a (the link size), and l_a/L_r is a weight representing the proportion contributed by link a to the overall route size. From Equation 3-7, it is observed that route size is inversely proportional to the degree of overlap, while the corresponding logarithm has a negative proportional relationship. Thus, the path size parameter is expected to be positive to allow for negative adjustment of route utility with increasing overlap.

The models accounting for overlap are generally expected to have better fit than the MNL model. Model fit is evaluated using the adjusted-rho square and the likelihood ratio tests (see Ben-Akiva and Lerman, 1985 for details).

3.6 Model results

This section presents the modelling results. We start by discussing the variable specification, followed by the model estimation and validation results.

3.6.1 Variable specification

The attributes available for possible inclusion in the model are summarised in Table 3-3. However, these could not all be specified together or in the same way for various reasons as explained below.

For travel time and cost, after initial tests using the linear specification, we used the log-transforms of the variables to allow for utility damping with respect to increasing time and cost (see Daly, 2010 for details). Various interactions of these variables with others (such as surface type) were tested, however, this led to correlation problems, and hence generic variables were specified.

The urban developments along the alternative routes were incorporated in terms of the average distance between towns rather than the number of towns to avoid situations where longer routes also have more towns. Again, this was specified using the log-transform of the variable for similar reasons as the travel time and cost. This being a largely rural road network with no traffic signals, the average distance between towns is the only variable we could use to capture traffic flow interruptions.

An attempt was made to incorporate scenic beauty into the model using either the length or the proportion of route length traversed through nature reserves, however, we could not achieve intuitive model results potentially due to our lack of detailed knowledge about the characteristics of these reserves, and the security levels of the corresponding routes. The final systematic utility specification is as follows;

$$V_{nr} = \beta_{l-cost} \ln(C_{nr}) + \beta_{l-time} \ln(T_{nr}) + \beta_{l-dtown} \ln(Dt_{nr}) \qquad (3\text{-}8)$$

Where C_{nr}, T_{nr}, and Dt_{nr} give the travel cost, the travel time, and the average distance between towns respectively of route r for individual n, and the βs are the corresponding model parameters to be estimated.

3.6.2 Estimation results

We present the of the MNL model, the c-logit model, and the path size logit based on the full sample in Table 3-6 for comparison purposes, where it is observed that most of the parameter estimates are statistically significant at the 95% level of confidence.

3.6.2.1 Route variables

The parameter signs for the travel cost variable are consistent with a priori expectations in each of the three models. In general, an increase in the cost of an alternative is expected to have a negative impact on its utility, hence the negative parameter sign. The same explanation holds for the travel time parameters as individuals generally prefer shorter travel times. On the other hand, the average distance between towns has a positive parameter sign. As earlier mentioned, this variable gives an indication of the amount of uninterrupted flow. Although no traffic congestion problems have been reported in these towns, traffic generally slows down due to speed control measures leading to delays. An increase in the average distance between towns therefore indicates more uninterrupted flow, hence the positive parameter sign.

Table 3-6 Estimation results

Variable	MNL model		C-logit model		Path size logit model	
	Parameter	**t-stat**	**Parameter**	**t-stat**	**Parameter**	**t-stat**
Route variables						
Natural log of travel cost (US Dollars)	-4.2221	-11.60	-4.1644	-11.50	-4.4117	-18.99
Natural log of travel time (Hours)	-1.6668	-13.78	-1.5987	-12.62	-1.1018	-17.87
Natural log of av. dist btn towns (Km)	0.5081	2.54	0.1705	1.10	1.5264	2.00
Commonality factor						
Beta			-0.0063	-1.48		
Gamma			0.6563	2.59		
Path size						
Path size parameter					1.6068	7.27
Measures of fit in estimation						
No. of observations	9453		9453		9453	
No. of decision makers	6497		6497		6497	
LL(C)	-11,135.41		-11,135.41		-11,135.41	
LL(F)	-7,758.83		-7,752.91		-7,547.60	
Number of parameters	3		5		4	
ρ^2_{adj} w.r.t LL(C)	0.3030		0.3033		0.3218	
LR w.r.t LL(C)	6,753.17		6,764.99		7,175.63	
p-value of LR	0.0000		0.0000		0.0000	

3.6.2.2 Overlap correction parameters

For the c-logit model, it is observed that the beta parameter of the commonality term is negative while the gamma parameter is positive. Similarly, the path size parameter in the path size logit model has a positive parameter sign. These results are in line with behavioural expectations as discussed earlier under Equations 3-6 and 3-7, an indication that CDR data is able to capture the behaviour towards overlapping routes.

3.6.2.3 Model comparison

A comparison of the adjusted rho-square values in Table 3-6 shows that the models accounting for overlap (i.e. the c-logit and the path size logit models) perform better than the MNL model. This is as expected given that the national road network of Senegal is highly overlapping (see Figure 3-1 and discussion under Figure 3-5). It is also worth noting that the path size logit model outperforms the c-logit model because the behavioural underpinning of the systematic utility adjustment process in the path size logit model is stronger than that in the c-logit model (Ramming, 2002).

The statistical significance of the improvements associated with accounting for overlap are evaluated using the likelihood ratios of the c-logit and the path size logit models with respect to the MNL model (see Table 3-7).

Table 3-7 Statistical comparison of the models

MNL formulation	**C-logit formulation**			**Path size logit formulation**		
LL(F)	**LL(F)**	**LR w.r.t MNL model**	**p-value**	**LL(F)**	**LR w.r.t MNL model**	**p-value**
-7758.83	-7752.91	11.83	0.0027	-7547.60	422.46	0.0000

From Table 3-7, it is noted that the p-values for the c-logit and the path size logit models are all less than 0.01, an indication that accounting for overlap has a statistically significant effect (at the 99% confidence level) beyond the improvements contributed by the additional degrees of freedom resulting from the extra parameters (see Ben-Akiva and Lerman, 1985 for details).

3.6.2.4 Policy insights

This section highlights the policy implications of the reported results in terms of the value of travel time (VTT). This metric quantifies the benefits derived from reduced travel time in monetary terms, and is useful in transportation cost-benefit analysis (Mackie et al., 2001). The value of travel time is calculated by taking the ratio of the partial derivatives of the systematic utility function (V) with respect to the travel time (T_{nr}) and cost (C_{nr}) as follows;

$$VTT = \frac{\partial V_{nr}/\partial T_{nr}}{\partial V_{nr}/\partial C_{nr}} = \frac{\beta_{l-time}}{\beta_{l-cost}}\frac{C_{nr}}{T_{nr}} \quad (3\text{-}9)$$

We computed the average VTTs for each model using the estimation data, and compared the values with those derived from other studies as well as other relevant statistics as summarised in Table 3-8.

Table 3-8 Comparison of the VTT estimates with other sources

Model	Values in USD/hr and 2013 prices			
	VTT current study	**VTT Teye et al. (2017) meta-analysis**	**Dakar–Diamniadio road toll (Gainer and Chan, 2016)**	**Median hourly wage (Tijdens et al., 2012)**
MNL	1.0822			
C-logit model	1.0524	4.3213	2.3411	0.6767
Path size logit model	0.6846			

In the Africa-wide meta-analysis by Teye et al. (2017), VTT was estimated as a function of the GDP per capita. However, the reported mean value (4.3213 USD/hr) seems high when compared to the toll being charged on the new Dakar–Diamniadio toll highway for a time saving of one hour (2.3411 USD/hr), a value that was highly criticised by the Senegalese media as being extremely high (Gainer and Chan, 2016). Although the median hourly wages do not necessarily translate into the value of travel time, they give a good indication of the range in which these values should fall, and as observed in Table 3-8, the average VTT estimate for the path size logit model is very close to the Senegalese median hourly wage. We consider this VTT estimate to be more reasonable for Senegal.

3.6.3 Validation results

The models based on the full sample provide intuitive results in terms of the parameter signs and the relative model performance. To assess the stability and predictive performance of the models, the dataset was randomly split into five parts at the individual level. Five rolling subsets, each comprising of 80% of the users were generated for model estimation purposes. For each of these, a complementary subset comprising of 20% of the users was generated for validation purposes. The models were re-estimated on each of the 80% subsets, and the parameter estimates applied to the corresponding 20% hold-out subsets to estimate the predictive measures of fit. Table 3-9 presents the summary outputs from this process.

The general interpretation of the parameter signs and the relative model performance in each of the 80% estimation subsets remained the same as in the full sample, an indication that the data is representative. A comparison of the measures of fit in estimation and validation shows that there is no significant loss in model fit, an indication that the performance of the models during estimation is not due to overfitting, rather it is due to the strong explanatory power of the variables.

A comparison of the predictive measures of fit shows that the relative model performance during estimation is mirrored on the holdout samples, with the path size logit model still giving the best model performance due to its behavioural superiority.

It would have been interesting to further validate the above results with outputs of route choice models based on traditional data or GSM data, but this was not possible due to lack of data in Senegal.

Table 3-9 Validation results

Subset	MNL model		C-logit model		Path size logit model	
	LL(F)	**Adjusted rho-square**	**LL(F)**	**Adjusted rho-square**	**LL(F)**	**Adjusted rho-square**
Estimation subsets (comprising of 80% of the users)						
Subset 1	-6201.66	0.3179	-6196.67	0.3182	-6045.15	0.3350
Subset 2	-6112.08	0.2999	-6106.49	0.3003	-5952.67	0.3180
Subset 3	-6306.93	0.2952	-6297.40	0.2961	-6138.29	0.3140
Subset 4	-6272.84	0.2993	-6263.20	0.3001	-6125.94	0.3156
Subset 5	-6184.90	0.2970	-6173.96	0.2980	-6018.68	0.3157
Validation subsets (comprising of 20% of the users)						
Subset 1	-1573.97	0.2267	-1569.71	0.2278	-1531.79	0.2469
Subset 2	-1661.15	0.3069	-1657.24	0.3077	-1621.68	0.3229
Subset 3	-1466.52	0.3265	-1466.01	0.3259	-1436.50	0.3398
Subset 4	-1502.27	0.3093	-1502.25	0.3084	-1450.65	0.3325
Subset 5	-1592.25	0.3165	-1590.68	0.3163	-1558.17	0.3306

3.7 Summary and conclusions

This paper has successfully demonstrated the potential of CDR data to capture rational route choice behaviour for long-distance inter-regional O-D pairs. The broad choice framework was used to leverage the limitations of CDR data where unique route choices could not be observed for some users, and only the broad sub-groups of the possible routes were identifiable. This study is unique in the sense that it adapts the broad choice framework to the context of route choice modelling using noisy CDR data.

An examination of the parameter signs shows that CDR data is able to capture the expected sensitivities towards particular route attributes. A review of different models accounting for overlap was conducted, and among these, the c-logit and the path size logit models were considered. A comparison of these models against the multinomial logit model (which does not account for overlap) showed significant improvements in model fit, with the path size logit model giving the best performance. The validation runs based on the 20% holdout samples largely showed the same advantages in prediction, especially for the path size logit model. These results show that CDR is able to capture the expected behaviour towards overlapping routes.

This study is timely as it extends the application of CDR data beyond travel pattern visualisation to econometric modelling of route choice. The proposed framework can help in the assessment of different policy implications at a low cost compared to traditional approaches, which involve expensive data collection. For example, the models developed in this study can be used to reliably estimate the value of travel time (VTT) as we have demonstrated. This study is thus beneficial to developing countries where budget constraints on transport studies are common and traditional data for transport studies is scarce.

We conclude that the study findings serve as a proof-of-concept that CDR data can be successfully used to model route choice behaviour for long-distance inter-regional trips, where there is a strong possibility that a user will use his/her phone during the trip, thereby enabling the capture of their partial trajectories needed for route identification. It may be noted that with the increasing trend of mobile internet usage (Gerpott and Thomas, 2014), the temporal resolution of CDR locations is likely to improve significantly in the near future, and this could make CDR data suitable for evaluating route choice behaviour for short trips.

A comparison of the study findings with those based on traditional data from Senegal would have been insightful, however, this was not possible due to data unavailability. Investigating the performance of the proposed approach in urban or intra-city scenarios would be an interesting direction for future research.

Acknowledgements

We would like to thank the Economic and Social Research Council (ESRC) of the UK and Institute for Transport Studies, University of Leeds for funding this research. The authors also acknowledge the financial support by the European Research Council through the consolidator grant 615596-DECISIONS. The research in this paper used the Data for Development (D4D) Challenge dataset from Senegal made available by Sonatel Group and Orange Group.

References

ADF 2011. Senegal Road Maintenance Project (PER): Project Completion Report (PCR). Tunis: African Development Fund.

Agard, B., Morency, C. & Trépanier, M. 2006. Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes,* 39**,** 399-404.

Ageroute Senegal. 2017. *Programmes & Projets* [Online]. Ageroute Senegal. Available: http://www.ageroute.sn/index.php/rapports-d-activités-annuels/cat_view.html [Accessed 12 April 2017].

ANSD 2016. Rapport projection de la population du Senegal. Dakar, Senegal: Agence Nationale de la Statistique et de la Démographie.

ANSD. 2017. *Situation Economique et Sociale* [Online]. Agence Nationale de la Statistique et de la Démographie. Available: http://www.ansd.sn/index.php?option=com_sess&view=sess&Itemid=398 [Accessed 11 April 2017].

ArcGIS. 2013. *Senegal Roads* [Online]. Available: https://services1.arcgis.com/4AWkjqgSzd8pqxQA/arcgis/rest/services/Senegal_Roads/FeatureServer/0 [Accessed 01 March 2017].

Autodesk 2017. AutoCAD Civil 3D. 2017 ed. California, United States of America: Autodesk Inc.

Azevedo, J., Costa, M. E. O. S., Madeira, J. J. E. S. & Martins, E. Q. V. 1993. An algorithm for the ranking of shortest paths. *European Journal of Operational Research,* 69**,** 97-106.

Bellman, R. & Kalaba, R. 1960. On k th best policies. *Journal of the Society for Industrial and Applied Mathematics,* 8**,** 582-588.

Ben-Akiva, M., Bergman, M., Daly, A. J. & Ramaswamy, R. Modeling inter-urban route choice behaviour. Proceedings of the 9th International Symposium on Transportation and Traffic Theory, 1984. VNU Science Press Utrecht, The Netherlands, 299-330.

Ben-Akiva, M. & Ramming, S. 1998. Lecture notes: Discrete choice models of traveler behavior in networks. *Prepared for Advanced Methods for Planning and Management of Transportation Networks. Capri, Italy,* 25.

Ben-Akiva, M. E. & Lerman, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

Ben-Elia, E. & Shiftan, Y. 2010. Which road do I take? A learning-based model of route-choice behavior with real-time information. *Transportation Research Part A: Policy and Practice,* 44**,** 249-264.

Bierlaire, M., Chen, J. & Newman, J. 2010. Modeling route choice behavior from smartphone GPS data.

Bierlaire, M., Chen, J. & Newman, J. 2013. A probabilistic map matching method for smartphone GPS data. *Transportation Research Part C: Emerging Technologies,* 26**,** 78-98.

Bierlaire, M. & Frejinger, E. 2008. Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies,* 16**,** 187-198.

Bitzios, D. & Ferreira, L. Factors affecting route choice of commercial vehicle drivers. PAPERS OF THE AUSTRALASIAN TRANSPORT RESEARCH FORUM, 1993.

Bolla, R., Davoli, F. & Giordano, F. Estimating road traffic parameters from mobile communications. Proceedings 7th World Congress on ITS, Turin, Italy, 2000.

Broach, J., Dill, J. & Gliebe, J. 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice,* 46**,** 1730-1740.

Cascetta, E., Nuzzolo, A., Russo, F. & Vitetta, A. A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. Proceedings of the 13th International Symposium on Transportation and Traffic Theory, 1996. Pergamon Lyon, France, 697-711.

Cascetta, E. & Papola, A. 2001. Random utility models with implicit availability/perception of choice alternatives for the simulation of travel demand. *Transportation Research Part C: Emerging Technologies,* 9**,** 249-263.

Chakirov, A. & Erath, A. 2012. Activity identification and primary location modelling based on smart card payment data for public transport.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Daganzo, C. F., Bouthelier, F. & Sheffi, Y. 1977. Multinomial probit and qualitative choice: A computationally efficient algorithm. *Transportation Science,* 11**,** 338-358.

Daly, A. 2010. Cost damping in travel demand models: Report of a study for the

Department for Transport. United Kingdom: RAND Corporation.

De La Barra, T., Perez, B. & Anez, J. Multidimensional path search and assignment. PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom, 1993.

De Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C. & Blondel, V. D. 2014. D4D-Senegal: the second mobile phone data for development challenge. *arXiv preprint arXiv:1407.4885*.

Deville, P., Song, C., Eagle, N., Blondel, V. D., Barabási, A.-L. & Wang, D. 2016. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences***,** 201525443.

Doyle, J., Hung, P., Kelly, D., Mcloone, S. F. & Farrell, R. 2011. Utilising mobile phone billing records for travel mode discovery.

Gainer, M. & Chan, S. 2016. A NEW ROUTE TO DEVELOPMENT: SENEGAL'S TOLL HIGHWAY PUBLIC-PRIVATE PARTNERSHIP, 2003 – 2013. New Jersey, USA: Innovations for Successful Socities, Priceton University.

Gerpott, T. J. & Thomas, S. 2014. Empirical research on mobile Internet usage: A meta-analysis of the literature. *Telecommunications Policy,* 38**,** 291-310.

Google Earth 7.1.8.3036. 2016. *Senegal roads, 14°24'13.70"N, 16°02'57.91"W, elevation 35m, imagery date January 2016* [Online]. Available: http://www.google.com/earth/index.html [Accessed 18 December 2017].

Google Maps. 2017a. *Google map directions* [Online]. Google. Available: https://www.google.co.uk/maps/dir [Accessed 13 April 2017].

Google Maps. 2017b. *Senegal nature reserves* [Online]. Google. Available: https://www.google.co.uk/maps/@15.1978209,-15.0824015,8.67z [Accessed 28 December 2017].

Görnerup, O. Scalable Mining of Common Routes in Mobile Communication Network Traffic Data. Pervasive, 2012. Springer, 99-106.

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly***,** 646-675.

GSM Association. 2017. *The Mobile Economy 2017* [Online]. Available: https://www.gsmaintelligence.com/research/?file=9e927fd6896724e7b26f33f61db5b9d5&download [Accessed 04 November 2017].

Hamerslag, R. 1981. Investigation into factors affecting the route choice in "Rijnstreek-West" with the aid of a disaggregate logit model. *Transportation,* 10**,** 373-391.

Hasan, S., Zhan, X. & Ukkusuri, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. Proceedings of the 2nd ACM SIGKDD international workshop on urban computing, 2013. ACM, 6.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P. & Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science,* 41**,** 260-271.

Hess, S., Quddus, M., Rieser-Schüssler, N. & Daly, A. 2015. Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review,* 77**,** 29-44.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C. & Pujolle, G. 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks,* 64**,** 296-307.

Imedia & Calao Production. 2013. *Le chemin de fer sénégalais* [Online]. AU-SENEGAL.COM. Available: http://www.au-senegal.com/le-chemin-de-fer,345?lang=fr [Accessed 06 August 2017].

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies,* 40**,** 63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A. & Willinger, W. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012. Acm, 239-252.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Kerali, H. G. R. 2000. *Overview of HDM-4,* Paris, The World Road Association (PIARC), Paris and The World Bank, Washington, DC.

Leontiadis, I., Lima, A., Kwak, H., Stanojevic, R., Wetherall, D. & Papagiannaki, K. From cells to streets: Estimating mobile paths with cellular-side data. Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies, 2014. ACM, 121-132.

Li, L., Wang, S. & Wang, F.-Y. 2018. An Analysis of Taxi Driver's Route Choice Behavior Using the Trace Records. *IEEE Transactions on Computational Social Systems,* 5**,** 576-582.

Logistics Cluster. 2013. *2.3 Senegal Road Assessment* [Online]. Logistics Cluster and World Food Programme. Available: http://dlca.logcluster.org/display/public/DLCA/2.3+Senegal+Road+Assessment [Accessed 11 May 2017].

Mackie, P., Jara-Dıaz, S. & Fowkes, A. 2001. The value of travel time savings in evaluation. *Transportation Research Part E: Logistics and Transportation Review,* 37**,** 91-106.

Mai, T., Fosgerau, M. & Frejinger, E. 2015. A nested recursive logit model for route choice analysis. *Transportation Research Part B: Methodological,* 75**,** 100-112.

Manski, C. F. 1977. The structure of random utility models. *Theory and decision,* 8**,** 229-254.

Marschak, J. 1960. Binary Choice Constraints on Random Utility Indications. *In:* ARROW, K. (ed.) *Stanford Symposium on Mathematical Methods in the Social Science.* Stanford, California: Stanford University Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics***,** 105-142.

Nie, J., Zhang, J., Zhong, G. & Hu, Y. 2015. A Novel Approach to Road Matching Based on Cell Phone Handover. *CICTP 2015.*

Nielsen, O. A. 2000. A stochastic transit assignment model considering differences in passengers utility functions. *Transportation Research Part B: Methodological,* 34**,** 377-402.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N. & Maurer, P. 2014. Supporting large-scale travel surveys with smartphones–a practical approach. *Transportation Research Part C: Emerging Technologies,* 43**,** 212-221.

Nitsche, P., Widhalm, P., Breuss, S. & Maurer, P. 2012. A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys. *Procedia-Social and Behavioral Sciences,* 48**,** 1033-1046.

Papinski, D., Scott, D. M. & Doherty, S. T. 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation research part F: traffic psychology and behaviour,* 12**,** 347-358.

Prato, C. & Bekhor, S. 2006. Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board***,** 19-28.

Prato, C. G. 2009. Route choice modeling: past, present and future research directions. *Journal of choice modelling,* 2**,** 65-100.

Qu, Y., Gong, H. & Wang, P. Transportation mode split with mobile phone data. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 285-289.

Ramming, M. S. 2002. *Network knowledge and route choice.* Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.

Reddy, S., Burke, J., Estrin, D., Hansen, M. & Srivastava, M. Determining transportation mode on mobile phones. Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on, 2008. IEEE, 25-28.

Rouphail, N. M., Ranjithan, S. R., El Dessouki, W., Smith, T. & Brill, E. D. A decision support system for dynamic pre-trip route planning. Applications of Advanced Technologies in Transportation Engineering, 1995. ASCE, 325-329.

Saravanan, M., Pravinth, S. V. & Holla, P. Route detection and mobility based clustering. Internet Multimedia Systems Architecture and Application (IMSAA), IEEE 5th International Conference, 2011. IEEE, 1-7.

Sayers, M. W., Gillespie, T. D. & Queiroz, C. a. V. 1986. The international road roughness experiment: establishing correlation and a calibration standard for measurements. *World Bank Technical Paper No. 45*.

Schlaich, J. 2010. Analyzing route choice behavior with mobile phone trajectories. *Transportation Research Record: Journal of the Transportation Research Board***,** 78-85.

Schlaich, J., Otterstätter, T. & Friedrich, M. Generating trajectories from mobile phone data. Proceedings of the 89th annual meeting compendium of papers, transportation research board of the national academies, 2010.

Sheffi, Y. & Powell, W. B. 1982. An algorithm for the equilibrium assignment problem with random link times. *Networks,* 12**,** 191-207.

Shier, D. R. 1979. On algorithms for finding the k shortest paths in a network. *Networks,* 9**,** 195-214.

Song, C., Koren, T., Wang, P. & Barabási, A.-L. 2010. Modelling the scaling properties of human mobility. *Nature Physics,* 6**,** 818-823.

Tettamanti, T., Demeter, H. & Varga, I. 2012. Route choice estimation based on cellular signaling data. *Acta Polytechnica Hungarica,* 9**,** 207-220.

Teye, C., Davidson, P., Porter, H. & Bell, M. G. H. 2017. Meta-analysis on the value of travel time savings in Africa. *International Choice Modelling Conference 2017.* Cape Town, South Africa.

Tijdens, K., Besamusca, J., Kane, A. & Tingum, E. N. 2012. Wages in Senegal - WageIndicator survey 2012. Amsterdam, The Netherlands: Wage Indicator Foundation.

Vovsha, P. & Bekhor, S. 1998. Link-nested logit model of route choice: overcoming route overlapping problem. *Transportation Research Record: Journal of the Transportation Research Board***,** 133-142.

Vrtic, M., Schuessler, N., Erath, A., Axhausen, K., Frejinger, E., Stojanovic, J., Bierlaire, M., Rudel, R. & Maggi, R. 2006. Including travelling costs in the modelling of mobility behaviour. Final report for SVI research program Mobility Pricing: Project B1, on behalf of the Swiss Federal Department of the Environment, Transport, Energy and Communications. IVT ETH Zurich, ROSO EPF Lausanne and USI Lugano.

Wang, H., Calabrese, F., Di Lorenzo, G. & Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, 2010. IEEE, 318-323.

Watanatada, T., Harral, C., Paterson, W., Dhareshwar, A., Bhandari, A. & Tsunokawa, K. 1987. The Highway Design and Maintenance Standards Model. *Volume 1 Description of the HDM-III Model.* Baltimore and London: The John Hopkins University Press.

White, J. & Wells, I. Extracting Origin Destination Information from Mobile Phone Data. Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), March 2002 London. IET, pp. 30 - 34.

WHO. 2016. *Global Health Observatory data repository: Maximum speed limits Data by country* [Online]. World Health Organisation. Available: http://apps.who.int/gho/data/view.main.51421 [Accessed 03 January 2018].

Wong, T. C. J. 2015. *Econometric Models in Transportation.* Ph.D. Thesis, University of California, Irvine.

Wootton, H., Ness, M. & Burton, R. 1981. Improved direction signs and the benefits for road users. *Traffic Engineering & Control*, 22.

World Bank 2004. Performance and impact indicators for transport in Senegal: Detailed statistics - June 2004. World Bank.

World Bank. 2017a. *Inflation, consumer prices (annual %) - Senegal* [Online]. The World Bank Group. Available: https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=SN [Accessed 02 Januray 2018].

World Bank. 2017b. *Lending interest rate (%) - Senegal* [Online]. The World Bank Group. Available: https://data.worldbank.org/indicator/FR.INR.LEND?locations=SN [Accessed 02 Januray 2018].

World Bank. 2017c. *Official exchange rate (LCU per US$, period average) - Senegal* [Online]. The World Bank Group. Available: https://data.worldbank.org/indicator/PA.NUS.FCRF?locations=SN [Accessed 02 Januray 2018].

Worldatlas. 2017. *Senegal Facts* [Online]. Worldatlas. Available: https://www.worldatlas.com/webimage/countrys/africa/senegal/snfacts.htm [Accessed 22 November 2017].

Zhang, L. & Levinson, D. 2008. Determinants of route choice and value of traveler information: a field experiment. *Transportation Research Record: Journal of the Transportation Research Board***,** 81-92.

Chapter 4

Modelling departure time choice using mobile phone data

Andrew Bwambale[*], Charisma F. Choudhury[*], Stephane Hess[*]

Abstract

The rapid growth in passive mobility tracking technologies has led to departure time choice studies based on GPS data in recent years. GPS data is however still expensive to collect and affected by technical issues like signal losses and battery depletion which create gaps in the data. On the other hand, the rapid growth in mobile phone penetration rates has led to the emergence of alternative passive mobility datasets such as Global System for Mobile communication (GSM) data. GSM data covers much wider proportions of the population and can be used to infer departure time information. This motivates this research where we rigorously compare the strengths and weaknesses of real-world GSM and GPS data to investigate their potential use for modelling departure time choice. We describe practical approaches to extract relevant information from the passive datasets and propose a modelling framework that accounts for the fact that the desired departure times are unobserved. We assume that the preferred departure times vary randomly across the users and apply the mixed logit framework to jointly estimate the distribution parameters of the preferred departure times and the sensitivities to schedule delay. We find that fewer time gaps in the GSM data lead to more reliable model results when compared against those based on GPS data, despite the higher location accuracy of the latter. This is also supported by the comparison of the valuation metrics derived from both models, where those obtained from GSM data are found to be closer to those based on traditional data sources.

Keywords: Time of travel, GSM data, GPS data, schedule delay, time valuation

[*] Choice Modelling Centre, Institute for Transport Studies, University of Leeds (UK)

4.1 Introduction

The modelling of time-of-travel choices has over the years emerged as an important and challenging issue worth consideration under travel demand management through policy measures such as congestion pricing and flexible working hours (Hess et al., 2007b). Time-of-travel choices are principally a trade-off between enduring longer travel times during peak demand periods to ensure punctual arrivals at the target destinations versus avoiding the peak periods and opting for earlier or later arrivals to reduce the travel times (Small, 1982), although in a toll road setting there may be additional differences in toll in the peak.

In most practical applications, time-of-travel choice problems have been expressed as scenarios where an individual is faced with a finite number of discrete departure time periods and chooses the alternative with the highest utility (Cosslett, 1977, Small, 1982). Departure time choice models, which are basic functions of the factors affecting departure time decisions are thus important tools for predicting travel demand and evaluating alternative measures for managing this demand.

Departure time choice models have largely been developed using traditional stated preference datasets (e.g. Hess et al., 2007b, Hess et al., 2005, De Jong et al., 2003, Daly et al., 1990, Bates et al., 1990) and revealed preference datasets (e.g. Bhat, 1998a, Bhat, 1998b, Small, 1987, Small, 1982, Abkowitz, 1981). The former are prone to hypothetical bias and behavioural incongruence while the latter are generally expensive to obtain, prone to reporting errors, and typically involve small samples. This problem is particularly common in developing countries, where stringent budget constraints for transport studies act as a barrier for large-scale data collection. Another reason for the use of stated preference data has been that the correlations inherent in revealed preference data make it difficult to capture the trade-offs between changes in departure time and other variables.

The last few decades have been characterised by rapid growth in technologies that enable the passive collection of individual mobility trajectories. This has led to a few departure time choice studies based on GPS data from smartphones (e.g. Peer et al., 2013). Although the use of smartphone apps has reduced costs, such studies remain expensive and thus usually involve small samples. Besides, enabling GPS often drains smartphone batteries and as a result, this functionality is often disabled by participants. Furthermore, GPS data is affected by technical issues such as signal losses in urban canyons, buildings, tunnels, and public transport vehicles such as buses and trains (Gong et al., 2012, Chen et al., 2010). This reduces the spatial and temporal coverage of the data and can make it difficult to capture the full set of trips made by individual travellers.

However, the rapid growth in mobile phone penetration rates worldwide (GSM Association, 2017) has led to the emergence of network-generated passive mobility datasets such as Call Detail Records (CDRs)[7] and Global System for Mobile communication (GSM)[8] data. These datasets can anonymously cover much wider proportions of the population using their current mobile handsets, without additional expenses such as recruiting of respondents, procuring of smartphones, and additional battery drainage issues. Such mobile phone

[7] CDR data typically consists of the time stamped locations of the responding tower that handles a call/text/web access request from a user as well as the details of the request (type, sender/receiver, etc.).

[8] GSM data reports the IDs of all the GSM cells traversed by an active mobile phone (i.e. a phone-set with a valid sim that is switched on) at regular time intervals.

datasets have been successfully used in various transportation planning applications (Çolak et al., 2015, Iqbal et al., 2014, Jiang et al., 2013, Isaacman et al., 2012, Schlaich, 2010). However, a review of the literature shows that there is no study using such data to model departure time choice decisions. This motivates this research where we investigate the potential of GSM data for departure time choice modelling. It may be noted that GSM data is deemed to be more appropriate for capturing departure time choices due to its semi-continuous nature as opposed to CDR location data, which is typically discontinuous.

Since GSM data generation only requires the users' mobile phones to be active, the regular location area updates by the network operator make it possible to capture most of the trips made. However, it is important to highlight the limitations of GSM data in the context of departure time analysis. The coarse location resolution of GSM data makes it impossible to capture intra-cell movements as well as the actual arrival or departure times from points within the cells. Instead, it is only possible to observe the cell boundary crossing times, especially where the GSM cells are recorded at short time intervals (e.g. 60 seconds in this study). It is worth noting that the differences between the actual departure and the (post-departure) cell boundary crossing times as well as the differences between the actual arrival and the (pre-arrival) cell boundary crossing times reduce as the GSM cell sizes become smaller. This is the case for most metropolitan areas where GSM cellular networks are dense, with small cell sizes that can go as low as 100 metres (e.g. De Groote, 2005). This implies that the cell boundary crossing times would still be within minutes from the actual departure or arrival times.

The above points motivate us to systematically compare the strengths and weaknesses of GPS versus GSM data in the context of departure time choice modelling as this could inform policy measures related to big data adoption for transport studies. We use the Nokia Mobile Data Challenge (MDC) dataset (Laurila et al., 2012, Kiukkonen et al., 2010) to critically compare the two data types and extract information for departure time analysis. Departure time choice models are then developed using advanced discrete choice modelling techniques. We focus on modelling departure time choices during peak periods as these are most critical in transport planning and operation. The study also proposes a theoretical approach for dealing with the absence of information on the desired times of travel in passively collected data. The proposed approach is unique in that it allows us to understand the sensitivities as well as the valuations attached to schedule delay despite the passive nature of the data. Furthermore, we propose a practical approach for imputing missing travel time data for some of the time intervals in the analysis period.

The remainder of the paper is arranged as follows; section 4.2 presents a brief review of relevant literature, section 4.3 describes the data used for this study and the associated challenges, section 4.4 presents the modelling framework, section 4.5 presents the model results, while section 4.6 presents the summary and conclusions of the study.

4.2 Literature review

Departure time choice decisions generally involve a trade-off between the travel time and the schedule delay associated with a given time period. However, estimating the schedule delay requires knowledge of the desired times of travel. Most stated preference datasets for departure time choice modelling collect information on the desired times of travel which makes it easy to estimate the schedule delay terms. However, this is not usually the case for revealed preference data, especially passively collected data such as mobile phone data. Peer

et al. (2013) is an exception where users were asked to report their desired times of travel. Previous studies have tried to address this issue in different contexts as summarised below.

Hess et al. (2007a) propose the use of time period specific constants to capture the aggregate scheduling preferences (among other effects) in the absence of the desired times of travel. However, Ben-Akiva and Abou-Zeid (2013) argue that the time period specific constants only capture the schedule delays if they are specified differently for each socio-economic group based on the assumption that individuals in the same socio-economic group have the same desired times of travel. This however results in the explosion of constants in the model specification, an issue that can be addressed with functional forms to approximate the alternative specific constants (Hess et al., 2005). However, another important point to highlight is that relying solely on constants to capture scheduling makes it difficult to understand the continuous sensitivity to delay.

On a different note, Koppelman et al. (2008) propose an approach where the schedule delay for a particular departure time period is estimated as the weighted mean of all the possible schedule delays with respect to the different time periods, where the weights are estimated from a time-of-day distribution of the observed departure times represented by a trigonometric function. However, a potential issue with this approach is that it assumes a strong correlation between the schedule delays and the observed time-of-day distributions, which may not be the case. A slightly related approach is proposed by Kristoffersson and Engelson (2018) who apply reverse engineering techniques that rely on a previously estimated departure time choice model to derive conditional departure time probabilities (given the preferred departure time), which are then combined with the observed departure time distributions for groups of O-D pairs to derive the weights for each preferred departure time period using ordinary least squares. However, a potential drawback with this approach is that previous models may be non-existent, and where they exist, there may be serious consequences with regard to model transferability.

Finally, Brey and Walker (2011) propose a hybrid choice framework in which the preferred times of travel are assumed to be latent and varying across individuals, and parameterise the probability density function as a mixture of normal distributions. However, in their framework, the latent preferred times of travel are explained using the trip and the travellers' characteristics, and are measured against the stated preferred times of travel obtained from a survey, which is not possible in this case study.

In this study, we propose a simple alternative approach which is described in Section 4.4 of this paper under the modelling framework.

4.3 Data

This study uses the Nokia Mobile Data Challenge (MDC) dataset collected as part of the Lausanne Data Collection Campaign (LDCC) between 2009 and 2011 (Laurila et al., 2012, Kiukkonen et al., 2010). The subsequent sections describe the study area, the mobile phone data, and the processes undertaken to extract relevant information for departure time analysis.

4.3.1 Study area

The main study area is Lausanne, located in southwestern Switzerland, however, the spatial coverage of the data covers the entire country.

Lausanne has a dense GSM cellular network with small cell sizes (see Schulz et al., 2012 for details). The small cell sizes make the area generally suitable for the current study as the actual departure or arrival times, which are unobservable for GSM data would still be within minutes of the observed cell boundary crossing times.

Another key aspect of Lausanne is that over 68% of the residents are working commuters, and over 90% of these use motorised transport modes, which are usually affected by peak period delays e.g. due to traffic congestion (ThemaKart, 2017). The travel times in Lausanne typically increase by 44% and 63% during the morning and the evening peak periods, respectively (TomTom, 2016).

4.3.2 Data description

The MDC dataset contains several types of records such as demographic data, GSM data, GPS data, call logs, and bluetooth data etc. However, this study only uses the demographics, the GSM, and the GPS data, which are described in the subsequent sections.

4.3.2.1 Demographic data

The MDC data is one of the few available mobile phone datasets with user demographic details, however, the sample size is small given that participation was voluntary. The available data comprises of 83 full-time workers. The other available demographics for each of these include the gender and the age-group as summarised in Table 4-1.

Table 4-1 Demographic data summary statistics

Characteristic	Description	Number	Proportion (%)
Gender	Female	23	27.71
	Male	60	72.29
Age-group	Under 28 years	24	28.92
	28 years and above	59	71.08

It is important to note that although demographic data is available in this case, such data is usually unavailable in most mobile phone datasets due to privacy reasons. Previous studies have focused on the subject of demographic prediction and how this can be incorporated into transport modelling frameworks (see Bwambale et al., 2017 for details). However, since this is not the main focus of this paper, we directly incorporate the reported demographics into the models.

4.3.2.2 GSM and GPS data

The GSM data reports all the GSM cells traversed by each user's mobile phone at an interval of approximately 60 seconds. The data contains approximately 24.8 million records generated by the full-time workers. Each record is described by a user ID, a unique internal ID of the GSM cell, the unix timestamp and time zone. Table 4-2 presents an excerpt of the GSM data.

Table 4-2 Excerpt of the GSM data

User ID	GSM Cell ID	Unix timestamp	Time zone
5451	686	1251762486	-7200
5451	686	1251762546	-7200
5451	1785	1251762606	-7200
5451	1785	1251762663	-7200

The GPS data (timestamped latitudes/longitudes) was collected concurrently with the GSM data using the users' smartphone GPS receivers, which allows for cross-comparison of the two datasets. Despite the higher time resolution of the GPS data versus the GSM data (i.e. 10 seconds versus 60 seconds), the GPS data contains only 5.2 million records.

GSM data was collected as long as the user's mobile phone was switched on albeit that signal losses were possible, while for GPS data, the facility needed to be enabled. The average data collection period per user was 278 days and 205 days for the GSM and the GPS data respectively. We use the concept of active time to refer to the time when a user's phone (or GPS service) is switched on, where this is assumed to be the case as long as the time interval between successive records does not exceed 10 minutes. For GSM data, the proportion of active time was on average 73.44% across users, compared to 5.00% for the GPS data. The corresponding median values and lower quartiles were 77.11% and 66.56% respectively for the GSM data compared to 4.37% and 3.16% respectively for the GPS data.

Time gaps in the GPS data may be caused by GPS disabling (e.g. due to battery issues) or signal loses in urban environments, buildings and tunnels (NCO, 2018, Gong et al., 2012, Chen et al., 2010). The methodology to extract meaningful information from both datasets is explained in the next section.

4.3.3 Data processing

The data processing methodology is presented in Figure 4-1. In this section, we briefly describe the key aspects of each major step.

4.3.3.1 Data preparation

a. GSM data

GSM data is noisy in nature as it sometimes contains cell jumps that do not represent actual movement. The noise is mainly caused by cell tower call balancing operations aimed at optimising the quality of calls, which makes mobile operators assign mobile phones to neighbouring cells even when these phones are not physically located within those cells (Çolak et al., 2015). To mitigate cell jumps, the ordered GSM cell sequence of each user was analysed to calculate the time periods between intermittent observations of the same cell, and those with time periods less than 10 minutes were treated as cell jumps, thus relabelling the cell observations between them accordingly (Iqbal et al., 2014). The cleaned GSM cell sequences were then analysed to extract the users' dwell locations as described in the next section.

b. GPS data

The technical issues associated with GPS data as highlighted in Section 4.3.2.2 may sometimes not lead to total signal loss, but rather, may lead to inaccurate GPS locations. Furthermore, GPS location accuracy is affected by factors such as the quality of the GPS

antenna in the smartphone, and the density of GPS satellites at the current location (NCO, 2018). Due to these factors, it is likely that different GPS points in the vicinity of one another could be linked to the same dwell location and this requires the application of spatial clustering techniques to identify the GPS point clusters. We conducted complete-linkage hierarchical clustering (Everitt et al., 2011, Murtagh, 1985) with a threshold distance of 300 meters as used in previous studies (Çolak et al., 2015, Jiang et al., 2013). This is detailed in Appendix A.

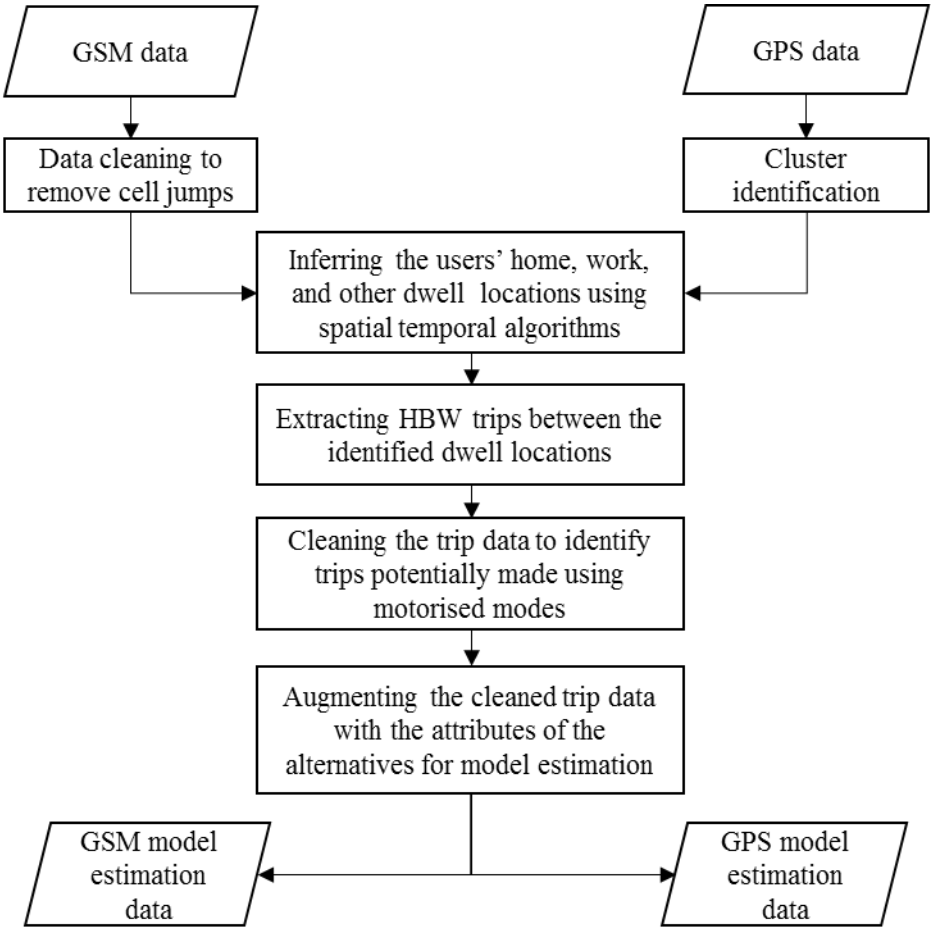


Figure 4-1 Summary of the data processing methodology

4.3.3.2 Identification of home and work locations

A home location was defined as the GSM cell or GPS point cluster in which a user was observed for the longest time between midnight and 6 am on a particular day, while a work location was defined as any location other than the home location in which a user (all of whom are workers) spent the longest time between 8 am and 5 pm on a particular working day. All the other GSM cells or GPS point clusters in which a user was seen to spend more than 10 minutes were described as 'other' dwell locations (Çolak et al., 2015, Jiang et al., 2013). We analysed each day separately to capture any possible changes in each user's home and work locations across the observation period.

An important point to note is that the original GSM data did not have the coordinates of the tower positions due to privacy reasons. The data only reports the IDs of the GSM cells without showing their positions as described in Section 4.3.2.2. Although the data alone is able to show the cell sequences and dwell times, it does not show the relative positions of these cells. To address this problem, the GPS points from all the users observed within 30 seconds of a GSM cell were extracted, and the mean latitudes and longitudes calculated for each cell. Although this was not possible for all the GSM cells, 62% of the inferred home cells, 70% of the inferred work cells, and 61% of the 'other' dwell locations were successfully matched. The data matching process was critical as it enabled the estimation of the distances and the travel speeds between the dwell cells, which we use to identify trips potentially made using motorised modes. It should be noted that under normal circumstances, GSM data should report the coordinates of the towers linked to the cells, in which case data matching would not be necessary.

4.3.3.3 Extracting HBW trips between the dwell cells

The focus of our analysis is home-based work (HBW) trips. Unlike direct trips, which have no en-route activity, trips with intermediate dwell locations (i.e. those labelled as 'others') could have en-route activities that last very long to the extent that such trips can no longer be categorised as clear HBW trips. In this study, we specified an upper limit of one hour[9] on the total duration across all the intermediate stops and included only the trips satisfying this criterion in our HBW model.

During the extraction of HBW trips, we checked whether each user's phone or GPS receiver was active both on departure and at arrival. For departure, we checked the time difference between the last observation in each departure dwell location and the first observation outside that location, while for arrival, we checked the time difference between the first observation in each arrival dwell location and the preceding observation outside that location. In this study, we specified a threshold of 2 minutes to ensure that we capture reasonably accurate trip start and end times without losing significant portions of the samples[10]. This, for example, helped us avoid situations where a user's phone was switched off during the trip, and switched back on several hours after arrival. The trips meeting all the above conditions were then taken through the subsequent stages as described in the next sections.

4.3.3.4 Cleaning the trip data to identify trips made using motorised modes

From a policy perspective, the focus is usually placed on motorised traffic, which is the main source of traffic congestion. However, one of the general limitations of mobile phone data is its anonymous nature, and therefore, the modes of transport used by the users are not known. A few previous studies have explored the possibility of detecting travel modes from mobile phone data (e.g. Qu et al., 2015, Doyle et al., 2011), however, as this is not the main focus of this study, we apply simple heuristics from the literature to infer the trips potentially

[9] The average commuting time in Lausanne is 36.5 minutes (ThemaKart, 2017) and specifying an upper limit of one hour ensures we do depart a lot from the mean value.

[10] 2 minutes corresponds to the 99th percentile time difference between subsequent GPS and GSM records in the full datasets excluding time-gaps above 10 minutes.

made using motorised modes. Observing a median speed above 15 kilometres per hour[11]. for a trip length above 5 kilometres is considered a good indicator that a user generally uses motorised transport for that trip chain (Hydén et al., 1999). Details of the applied heuristics are presented in Appendix B.

4.3.3.5 Augmenting the cleaned trip data with the attributes of the alternatives

Although it may seem convenient to assume that a user's choice set only comprised of the departure time intervals ever observed for the user across the different days in the sample, such an assumption is unrealistic since the failure to observe certain time intervals does not necessarily mean they were not considered. It seems more reasonable and safer to assume that all the departure time intervals were available and potentially considered. This implies a need to calculate the attributes for those time periods for which no actual trips were observed, a process described in this section.

The morning and the evening peak periods were divided into 15-minute intervals. The average travel times associated with each of these intervals were estimated for each of the user's trip chains using timestamps in their cleaned GSM and GPS data. The estimated travel times were then combined with time-period specific congestion factors to impute the travel times for the unobserved time intervals. The time-period specific congestion factors were estimated with the aid of the Google Maps direction tool, which predicts the average travel times between a given O-D pair at different departure or arrival times (Google Maps, 2018).

To reduce this task to manageable proportions, we divided Lausanne into 10 representative zones bounded by the major roads, thereby generating 90 O-D pairs as shown in Figure 4-2. For each O-D pair, we extracted the travel times associated with each 15-minute interval between 5 am and 11 am (for the home-to-work commute) and 3 pm to 9 pm (for the work-to-home commute). The average travel times for each interval across all the O-D pairs were then determined.

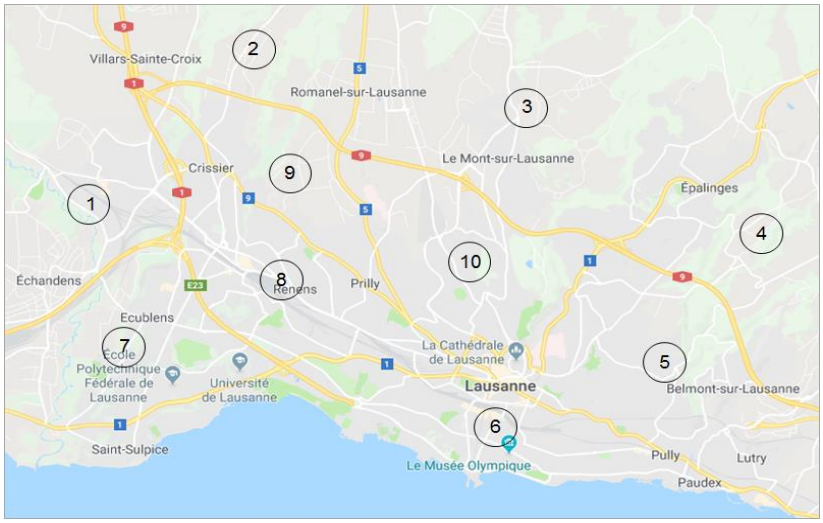


Figure 4-2 Sample zones for travel time analysis (Google Maps, 2018)

[11] A median speed of 15 km/h is not a very restrictive threshold as the average peak period speeds in Lausanne are close to this value (Google Maps, 2018), and less than 15% of the data was discarded by imposing this threshold.

For each analysis period, we computed the time-period specific congestion factors by first establishing the interval with the shortest average travel time, and calculating the ratios of the travel times for each interval versus the minimum travel time for the analysis period as illustrated in Figure 4-3.

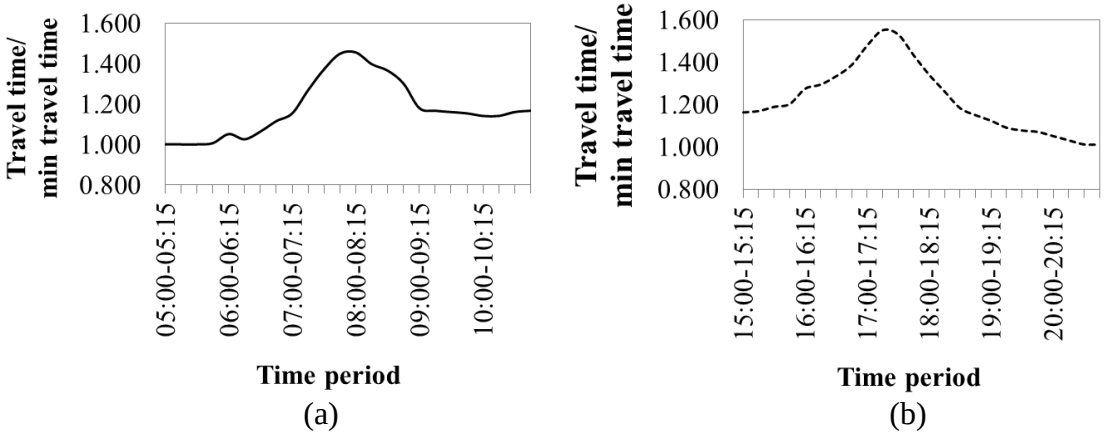


Figure 4-3 Travel time variation (a) Morning peak, (b) Evening peak

Given the typical ratios for each time interval, we used the observed average travel times for each user (i.e. those based on the cleaned GSM and GPS data) to estimate the minimum travel times for each of their trip chains, and applied the appropriate time-period specific congestion factors to impute the travel times for each time interval. We used the imputed travel times for both the chosen and the unchosen alternatives as this mitigates possible endogeneity bias which could arise from interrelationships with other underlying factors (Calastri et al., 2017, Sanko et al., 2014).

4.3.4 Comparison of the GPS and the GSM processed data

Due to differences in the temporal coverage of the GSM versus the GPS data (see Section 4.3.2.2), the inferred home and work locations were different for some users as illustrated in Figure 4-4.

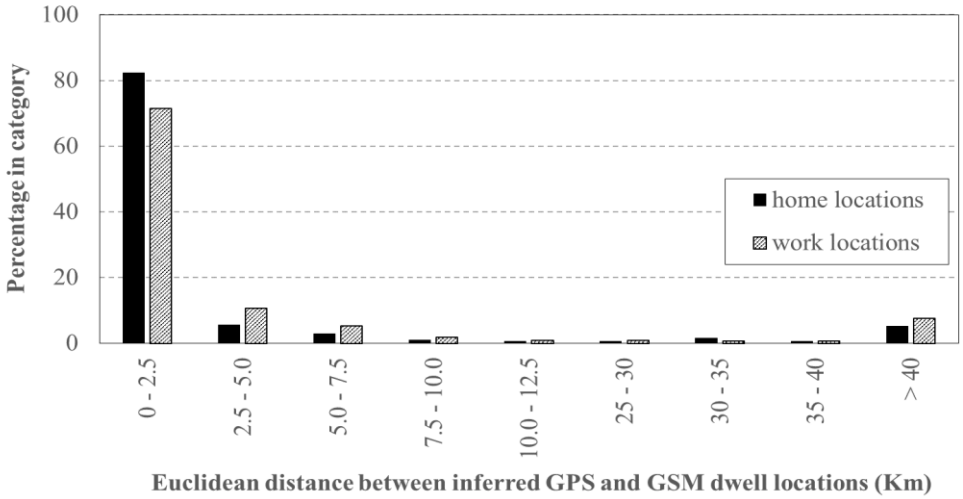


Figure 4-4 Differences between the GPS and GSM inferred home/ work locations

As observed, most of the inferred GPS and GSM home/work locations are within 5 kilometres of each other, an indicator that most belong to the same cell. However, we also have scenarios where the inferred dwell locations for the same day are over 40 kilometres apart. In such cases, GSM data, which has a higher temporal coverage, is expected to be more reliable. However, since our focus is on parameter comparison in the model development stage, we retain the dwell locations for each data type as extracted.

Furthermore, it is observed that the extracted departure time distributions are different across the two datasets, with the home-to-work commute having more pronounced differences (see Figure 4-5). This is because the time gaps in the night GPS traces are more than those in the daytime GPS traces, potentially because the users disable their GPS receivers more at night. This is probably the reason why the peak period for the home-to-work commute is not clearly defined. On the other hand, the peak periods for the GSM data are clearly observed for both the home-to-work, and the work-to-home commute. This is in line with the expected behaviour of full-time workers and is another indication of the reliability of GSM data.

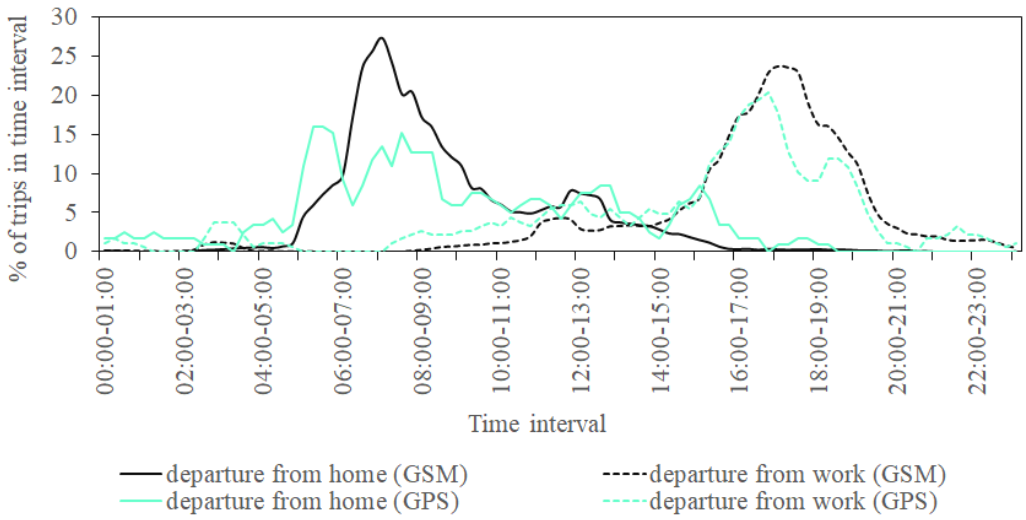


Figure 4-5 Commuter trip frequency distribution

4.4 Modelling framework

Our analysis is based on the random utility framework (Marschak, 1960), and theoretical insights from the scheduling model by Small (1982). Let U_{ntk} be the utility for individual n derived from departing in time period t in choice situation k. This can be expressed as;

$$U_{ntk} = V_{ntk} + \varepsilon_{ntk} \tag{4-1}$$

Where V_{ntk} and ε_{ntk} are the systematic and the random parts of utility, respectively. Based on scheduling theory, V_{ntk} is usually expressed as a function of the travel time, the corresponding schedule delays, and any other key attributes as follows;

$$V_{ntk} = \beta_{t-time}T_{ntk} + \beta_{l-dummy}L_{ntk} + E_{nt}(\beta_{e-time}SDE_{nt}) + L_{nt}(\beta_{l-time}SDL_{nt}) + \cdots \tag{4-2}$$

Where T_{ntk} is the travel time associated with time period t in choice situation k for individual n. E_{nt}, L_{nt}, SDE_{nt}, and SDL_{nt} are the earliness dummy, the lateness dummy, the

amount of earliness, and the amount of lateness associated with time period t for individual n in choice situation k. The β terms are the corresponding model parameters to be estimated. It may be noted that the earliness terms (E_{nt} and SDE_{nt}) and the lateness terms (L_{nt} and SDL_{nt}) are mutually exclusive.

Estimating the amount of earliness or lateness associated with a particular departure time period requires information on the desired times of travel, which is not available in this case as we are relying on passively collected datasets. However from a theoretical perspective, we know that every individual makes efforts to depart at his/her desired time to minimise scheduled delay. By re-writing Equation 4-2, this may be expressed as follows;

$$V_{ntk} = \beta_{t-time} T_{ntk} + \beta_{l-dummy} L_{nt} + E_{nt}(\beta_{e-time}[PDT_n - D_t]) + L_{nt}(\beta_{l-time}[D_t - PDT_n]) + \cdots \quad (4\text{-}3)$$

Where PDT_n is the preferred departure time for individual n, and D_t is the midpoint of departure time interval t in terms of the hours since midnight (e.g. for departure time interval 8:00 am – 8:15 am, D_t is 8.125 hours, which corresponds to 8:07:30 am).

In the absence of the preferred departure times of the users, it is reasonable to assume that these vary randomly across the individuals following a certain statistical distribution. The objective we are trying to pursue is to estimate the mean and standard deviation of this statistical distribution. The use of statistical distributions helps us to avoid the assumption that the preferred departure times follow the same trend as the observed departure times as used in Koppelman et al. (2008).

However, estimating the above specification presents serious identification and optimisation issues. This is because the earliness and lateness dummies depend on the preferred departure time, which is also being estimated at the same time. As the optimiser tries to find the preferred departure time, the dummies keep alternating between 0 and 1, thereby resulting in a function that is not continuously differentiable. This prompts us to deviate from Small's model by investigating alternative schedule delay functions that are first of all behaviourally intuitive, and continuously differentiable.

From a behavioural perspective, the schedule delay function needs to reflect reductions in the schedule disutility as the observed departure times approach the preferred departure times (from both the earliness and lateness sides), and must peak at points where the delay is zero. Furthermore, the function needs to have an indifference region around the preferred departure time to reflect the fact that the rate of increase in the schedule disutility is small around the preferred departure times, and increases as the observed departure times spread further away from the preferred departure time.

After testing various functional forms (e.g. the logistic and the parabolic functions), we selected the parabolic function, which gave consistent and intuitive results. The systematic utility is now expressed as follows;

$$V_{ntk} = \beta_{t-time} T_{ntk} + \alpha(PDT_n - D_t)^2 + \cdots \quad (4\text{-}4)$$

Where α is a parameter to be estimated, representing the sensitivity to delay. For the schedule function above to be behaviourally intuitive, the parameter α is expected to have

a negative sign. A potential issue with this function is that it does not capture the damping effect of schedule delay on marginal disutility as argued in previous studies (e.g. Koppelman et al., 2008), which calls for further research to address this issue.

Assuming the βs, the αs and PDT_n (for individual n) are known, and the random part of utility ε_{ntk} is independently and identically distributed across the choice situations, the alternatives, and the individuals, the departure time choice probability for individual n can be estimated using the multinomial logit (MNL) model (see McFadden, 1974 for details). However in this case, we have several choice situations for the same individual across different days, thus, we need to capture the panel effect while calculating the choice probabilities.

Let $P_{n,k}(t|\beta, \alpha, PDT_n)$ denote the logit probability that individual n chooses departure time period t in choice situation k, conditional on β, α and PDT_n. Furthermore, let $\hat{t}_{n,k}$ be the departure time chosen by individual n in choice situation k, such that $P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n)$ gives the logit probability of the observed choice for individual n in choice situation k, conditional on β, α and PDT_n. The logit probability of individual n's observed sequence of choices is;

$$P_n(\beta, \alpha, PDT_n) = \prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) = \prod_{k=1}^{K} \frac{\exp(V_{n\hat{t}k}|\beta, \alpha, PDT_n)}{\sum_{t^* \in C_n} \exp(V_{ntk^*}|\beta, \alpha, PDT_n)} \quad (4\text{-}5)$$

Where C_n is the choice set. It is important to note that for users with more than one trip chain, we compare the attributes of the same trip chain across the different time periods while computing the choice probabilities. That is, each trip chain represents a different choice scenario.

However as earlier mentioned, the preferred departure times PDT_n are not observed, and are assumed to vary randomly across individuals. Suppose PDT_n is independently and identically distributed over the individuals with density $f(PDT|\Omega)$, where Ω is a vector of the parameters of this distribution, such as the mean and standard deviation, this would result in the mixed multinomial logit (MMNL) model (McFadden and Train, 2000), and the mixed logit probability would be given by;

$$P_n(\beta, \alpha, \Omega) = \int_{PDT} \left[\prod_{k=1}^{K} P_{n,k}(\hat{t}_{n,k}|\beta, \alpha, PDT_n) \right] f(PDT|\Omega)\, dPDT \quad (4\text{-}6)$$

The integration over the density of PDT is done over all the individual's choices combined, since the same PDT applies to all the choice situations. The log-likelihood (LL) function for the observed choices is;

$$LL(\beta, \alpha, \Omega) =$$

$$\sum_{n=1}^{N} \ln\left(\int_{PDT} \left[\prod_{k=1}^{K} P_{n,k}\left(\hat{t}_{n,k} \middle| \beta, \alpha, PDT_n\right) \right] f(PDT|\Omega)\, dPDT \right) \quad (4\text{-}7)$$

An important consideration is the choice of distribution to be used. Due to our limited knowledge of the individuals' preferences, coupled with the fact that we have not conducted any surveys to determine the distribution of the preferred departure times, we assume a truncated normal distribution bounded between the limits of the analysis period (i.e. the morning or evening peak periods). Since the integral in Equation 4-7 has no closed form, it is estimated using simulation methods. The simulated log-likelihood (SLL) is expressed as follows;

$$SLL(\beta, \alpha, \Omega) = \sum_{n=1}^{N} \ln\left(\frac{1}{R} \sum_{r=1}^{R} \left[\prod_{k=1}^{K} P_{n,k}\left(\hat{t}_{n,k} \middle| \beta, \alpha, PDT_n\right) \right] \right) \quad (4\text{-}8)$$

The PDT distribution parameters (i.e. the mean and standard deviation) are estimated alongside the other model parameters by maximising the simulated log-likelihood using 300 Halton draws per user (Bhat, 2001). During parameter estimation, there may be a possibility of confounding between random PDT and random schedule delay sensitivity α, however, this is mitigated by applying the same parameter α to ${PDT_n}^2$, ${D_t}^2$, and $-2PDT_nD_t$ (see Equation 4-4).

4.5 Model results

This section discusses the process of variable specification, the estimation results, as well as the policy insights derived from the estimation results.

4.5.1 Variable specification

The variables available for possible inclusion in the departure time utility equation are; travel time, latent schedule delay, trip chain characteristics, and user demographics. However, each of these variables was defined in a particular way for different reasons as explained in the subsequent paragraphs.

For travel time, we tested the logarithmic specification to allow for damping effects (Daly, 2010) and found no gains in model fit compared to the linear specification. This could be attributed to the small ranges of travel time across the alternatives of each user. Therefore, we adopted a linear specification.

|The schedule delay function was entered into the model as specified in Equation 4-4. We investigated the possibility of different PDT distribution parameters for different demographic groups and could not obtain significant gains in model fit for either dataset.

The trip chain characteristics were incorporated into the model using the number of intermediate stops, and time-period specific parameters were specified to capture the differential impact on utility across the time periods. It may be noted that the duration at the intermediate stops is already incorporated into the travel time.

A number of interactions of the schedule delay and the travel time parameters with the user demographics were tested. For the GPS data, we could not obtain intuitive results for all the interactions tested due to the small sample size per demographic group in the final sample, so we specified generic parameters. On the other hand, for the GSM data, we successfully interacted the travel time and the schedule delay parameters with age-group alone and age-group by gender, respectively. The final systematic utility specifications for the GSM and the GPS data are given by Equations (4-9) and (4-10), respectively;

$$V_{ntk} = \beta_{time-age} T_{ntk} + \alpha_{del\_age\_gender} {SD_{nt}}^2 + \beta_{stops\_t} N_{stops} \quad (4\text{-}9)$$

$$V_{ntk} = \beta_{time} T_{ntk} + \alpha_{del} {SD_{nt}}^2 + \beta_{stops\_t} N_{stops} \quad (4\text{-}10)$$

Where $SD_{nt} = (PDT_n - D_t)$, and N_{stops} is the number of intermediate stops in the trip chain. The β and α parameters are to be estimated.

4.5.2 Estimation results

We present the estimation results for the home-to-work commute and the work-to-home commute models for both the GSM and the GPS data in Table 4-3 for comparison purposes. As observed, most of the parameter estimates are statistically significant at the 95% level of confidence. In the subsequent sections, we discuss each aspect of the results in details.

4.5.2.1 Distribution parameters for the departure time distribution

We specified a truncated normal distribution for the preferred departure time (for reasons explained in the paragraph after Equation 4-7). For this distribution, the estimated mean and standard deviation are those of the underlying normal distribution. To calculate the true means and standard deviations, the estimated parameters were adjusted as follows;

$$\mu = \hat{\mu} + \left[\frac{\phi(\mathrm{A}) - \phi(\mathrm{B})}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})} \right] \hat{\sigma} \quad (4\text{-}11)$$

$$\sigma = \hat{\sigma} \left[1 + \frac{\mathrm{A}\phi(\mathrm{A}) - \mathrm{B}\phi\mathrm{B}}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})} - \left(\frac{\phi(\mathrm{A}) - \phi(\mathrm{B})}{\Phi(\mathrm{B}) - \Phi(\mathrm{A})} \right)^2 \right]^{1/2} \quad (4\text{-}12)$$

Where, $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation respectively of the underlying normal distribution, μ and σ are the true mean and standard deviation respectively of the truncated distribution. $\mathrm{A} = (a - \hat{\mu})/\hat{\sigma}$, $\mathrm{B} = (b - \hat{\mu})/\hat{\sigma}$, where a and b are the lower and upper bounds respectively of the truncated distribution. For the home-to-work commute, these are set to 5 and 11, respectively, while for the work to home commute, these are 15 and 21 respectively.

From Table 4-3, it is observed that the mean preferred departure times for the home-to-work commute are 7.9742 and 8.0105 (approximately 8:00 am), while those for the work-to-home commute are 17.7240 (approximately 05:45 pm) and 17.2903 (approximately 05:15 pm) in the GSM and the GPS models, respectively. Although flexible working time is not unusual in Switzerland, normal business hours generally start between 08:00 am and 08:30 am and end between 05:00 pm and 06:30 pm (Switzerland Tourism, 2018), which is consistent with our findings. Furthermore, it is observed that the corresponding standard deviations are

slightly higher for the home-to-work commute when compared to the work-to-home commute. The lower amount of variation during the work-to-home commute is probably the reason behind the higher evening traffic congestion in Lausanne and other major Swiss cities (TomTom, 2016).

4.5.2.2 Sensitivity to schedule delay

Generally, individuals prefer to depart at particular times due to certain constraints at both the origin and the destination. Thus, any deviations from the desired times of travel are expected to cause disutility, hence the negative parameter signs for the schedule delay terms reported in Table 4-3.

For GSM data where we have different schedule delay parameters for different demographic groups, where we note that female workers are more sensitive to shifting departure time compared to male workers in the same age-group. This is the case during both the home-to-work and the work-to-home commute. The higher sensitivity of female workers is potentially attributed to the strictness in their schedule as a result of the need to balance family and professional life in the face of common views on traditional gender roles in Switzerland (Nguyen, 2018, The Economist, 2018).

Furthermore, it is observed that younger workers are more sensitive to schedule delay than older workers of the same gender during the home-to-work commute. This is expected as younger workers are more junior and typically have less flexibility (i.e. expected to report on time). However, the situation is different for the work-to-home commute. Here, it is observed that older female workers are more sensitive than young female workers. This again could be attributed to the levels of responsibility at home as older female workers are more likely to have families already. However, the lower sensitivity of older male workers in comparison with younger male workers is an interesting observation that needs to be investigated further.

For the GSM data, we also observe that the sensitivity to schedule delay is generally higher during the home-to-work commute when compared to the work-to-home commute. This is expected as late arrival on the home-to-work commute probably has more serious consequences than on the work-to-home commute. However, this was not captured in the GPS data due to differences in the sample composition resulted by the big time gaps in the GPS data.

4.5.2.3 Sensitivity to travel time

From Table 4-3, the parameter signs for the travel time variable are negative in both the home-to-work and the work-to-home commute models, which is consistent with a priori expectations. In general, time periods with higher travel times are less attractive, and prompt individuals to choose earlier or later time periods at the expense of increasing the schedule delays. Keeping all other things constant, it is observed that the sensitivity to travel time during the home-to-work commute is generally higher than that during the work-to-home commute in both the GSM and the GPS models. This implies that people are less willing to spend longer times in traffic during the home-to-work commute as opposed to the reverse direction, which could be attributed to the higher stakes attached to the home-to-work leg.

Table 4-3 Model estimation results

Variable	Home-to-work commute				Work-to-home commute			
	GSM data		GPS data		GSM data		GPS data	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
Travel time (hours)								
Workers < 28 years	-2.9700	-2.37	-0.9486	-1.05	-1.0674	-2.59	-0.2340	-0.32
Workers >= 28 years	-1.3266	-1.35			-1.0314	-2.15		
Schedule delay term (hours^2)								
Female workers < 28 years	-0.7682	-6.51	-0.2560	-1.88	-0.4412	-5.70	-0.3093	-6.13
Female workers >= 28 years	-0.6441	-5.34			-0.4785	-7.12		
Male workers < 28 years	-0.6841	-3.49			-0.4148	-4.46		
Male workers >= 28 years	-0.5468	-6.20			-0.3749	-6.78		
Preferred departure time distribution parameters								
$\hat{\mu}$	7.9652	67.55	8.0105	18.76	17.7240	170.59	17.2903	112.59
μ	7.9661		8.0081		17.7244		17.2903	
$\hat{\sigma}$	1.0029	7.85	1.5077	1.99	0.7465	11.98	0.4729	4.40
σ	0.9783		1.7503		0.5562		0.2236	
Number of stops (Time period specific parameters)*								
β_{stops_1}	-2.2939	-2.24	-1.1547	-0.43	1.8821	1.97	-2.4504	-2.17
β_{stops_2}	-0.8544	-1.80	-1.3607	-0.53	2.0265	2.18	-2.9766	-2.63
β_{stops_3}	-1.3742	-2.05	-1.5777	-0.66	1.8313	2.00	-3.4635	-3.05
β_{stops_4}	-0.3398	-0.81	16.7402	6.94	1.2878	1.39	11.7430	7.91
β_{stops_5}	-1.1779	-2.68	-1.9351	-0.91	1.7281	1.87	11.5697	8.53
β_{stops_6}	-1.1063	-2.76	-2.1946	-1.13	1.3622	1.44	12.1103	8.28

Table 4-3 cont'd

Variable	Home-to-work commute				Work-to-home commute			
	GSM data		GPS data		GSM data		GPS data	
	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat	Parameter	t-stat
β_{stops_7}	-1.2352	-3.28	-2.3215	-1.26	1.8133	1.95	-4.7535	-4.36
β_{stops_8}	-1.0663	-2.88	-2.3982	-1.43	1.5522	1.68	-4.8664	-4.55
β_{stops_9}	-1.0169	-3.17	15.1030	8.56	1.6290	1.77	11.1953	8.50
β_{stops_10}	-0.6124	-1.70	15.1026	11.49	1.4974	1.61	-4.7149	-4.67
β_{stops_11}	-1.2679	-3.30	-2.2488	-1.79	1.8324	1.99	-4.7034	-4.72
β_{stops_12}	-1.1857	-3.69	-2.1632	-1.87	1.3271	1.45	11.2286	8.07
β_{stops_13}	-0.9265	-2.95	-2.1864	-2.09	1.1121	1.19	-4.6527	-4.75
β_{stops_14}	-1.1458	-2.84	-2.3160	-2.67	0.8245	0.88	-4.4459	-4.61
β_{stops_15}	-1.1957	-2.61	-2.3777	-3.25	0.8900	0.94	-4.1544	-4.35
β_{stops_16}	-1.7638	-4.76	-2.4956	-4.26	0.3669	0.38	-3.7369	-4.05
β_{stops_17}	-1.1030	-2.80	-2.7512	-6.41	1.1553	1.26	-3.2570	-3.67
β_{stops_18}	-2.8825	-4.97	-2.7162	-8.04	1.2105	1.30	-2.7495	-3.25
β_{stops_19}	-1.0044	-3.49	-2.6429	-10.06	0.9961	1.03	-2.2151	-2.77
β_{stops_20}	-1.2905	-4.90	-2.5495	-13.12	0.9618	1.01	-1.7011	-2.28
β_{stops_21}	-1.2117	-4.28	-2.4537	-15.91	1.1151	1.21	-1.2485	-1.80
β_{stops_22}	-0.5033	-1.98	-2.3064	-17.13	0.7528	0.79	-0.8655	-1.35
β_{stops_23}	-0.6143	-2.71	-2.0963	-16.51	1.6520	1.89	-0.5606	-0.96

Table 4-3 cont'd

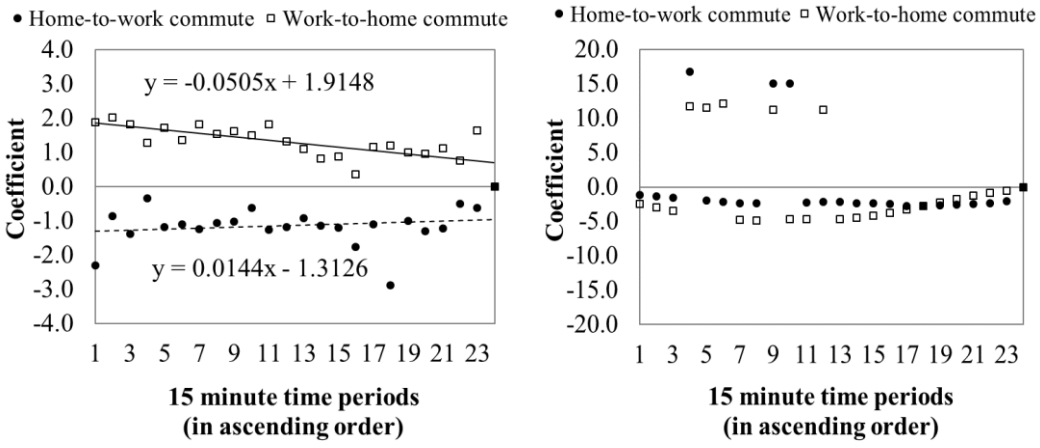
Variable	Home-to-work commute		Work-to-home commute	
	GSM data	**GPS data**	**GSM data**	**GPS data**
Measures of fit in estimation				
Number of observations	2043	69	2668	112
Number of decision makers	78	29	78	35
LL(C)	-6492.76	-219.29	-8479.05	-355.94
LL(F)	-5575.19	-203.50	-7644.19	-323.95
Number of parameters	31	27	31	27
ρ^2_{adj} w.r.t LL(C)	0.1365	-0.0511	0.0948	0.0140
LR w.r.t LL(C)	1835.15	31.58	1669.71	63.99
p-value of LR	0.0000	0.2480	0.0000	0.0001

* 1 refers to the first 15-minute interval in the period of analysis (i.e. 5:00 to 5:15 for morning home-to-work commute, and 15:00 to 15:15 for the evening work-to-home commute). The rest of the numbers refer to the subsequent 15-minute intervals in ascending order.

4.5.2.4 Time-period specific parameters related to the number of stops

Another important issue worth highlighting concerns the time-period specific parameters related to the number of stops. For easy parameter identification, we normalised to zero the effect linked to the last departure time interval of each analysis period (i.e. 10:45 am to 11:00 am for the morning home-to-work commute and 08:45 pm to 09:00 pm for the evening work-to-home commute). Thus, the reported parameters represent the differential impact on utility with respect to the reference time periods, and can either be positive or negative. These are reproduced in Figure 4-6 for easy visualisation.

The interpretation of these parameters is however difficult given that they probably incorporate some other unobserved factors associated with the different time periods. Nevertheless, for GSM data (Figure 4-6a), the trend of the parameters in the work-to-home model shows that if someone is to make a trip with more stops, they will find the earlier departure time periods more suitable, which is reasonable. However, the trend is different in the home-to-work model, where it is observed that the later departure time periods would be more suitable. This could be attributed to the less traffic during the inter-peak period and the opening times at the various stop locations. Of course, it is not clear if someone is choosing a specific departure time given the stops they plan to make, or if trips at specific departure times imply different stop patterns. For GPS data (Figure 4-6b), there is no clear trend in the parameters for both commute directions, which is attributed to the gaps in the data.



(a) GSM data (b) GPS data

Figure 4-6 Time period specific parameters for the number of stops

4.5.2.5 Overall model performance

From Table 4-3, it is observed that the adjusted-rho square values of the GPS models are smaller compared to those of the GSM models. While acknowledging that the models cannot be directly compared due to differences in the sample compositions, the overall poor performance of the GPS models is largely attributed to the small GPS sample sizes.

4.5.3 Policy insights

Travel time is usually at its worst when the schedule delay for most individuals is at the minimum. Therefore, people are generally faced with a trade-off between travel time and

schedule delay when choosing the most appropriate departure time periods. Thus, to gain better insights, it is critical to analyse the sensitivities to schedule delay versus travel time to obtain the values attached to schedule delay.

The time valuation of schedule delay ($TVSD$) represents the amount of delay an individual is willing to experience for a unit reduction in travel time by changing his/her departure schedule. This unitless metric is calculated as the ratio of the partial derivatives of the systematic utility with respect to schedule delay and travel time as follows;

- For the GSM data

$$TVSD_{age\_gender} = \frac{\partial V_{ntk}/\partial SD_{nt}}{\partial V_{ntk}/\partial T_{ntk}} = \frac{\alpha_{del\_age\_gender} * 2SD_{nt}}{\beta_{time-age}} \quad (4\text{-}13)$$

- For the GPS data

$$TVSD = \frac{\partial V_{ntk}/\partial SD_{nt}}{\partial V_{ntk}/\partial T_{ntk}} = \frac{\alpha_{del} * 2SD_{nt}}{\beta_{time}} \quad (4\text{-}14)$$

Since we used square-transformations, the time valuation of schedule delay depends on the amount of earliness/lateness of the individual as shown in Equations 4-13 and 4-14. We therefore used the estimation data (including the normal draws) to calculate the average values across individuals, which we report in Table 4-4.

Table 4-4 Time valuations of schedule delay

Commute direction	GSM data					GPS data
	Female worker < 28 years	Female worker >= 28 years	Male worker < 28 years	Male worker >= 28 years	Weighted mean	Generic
Home to work	1.2196	2.5116	0.9995	2.2031	2.0167	2.3435
Work to home	1.6729	1.8290	1.5570	1.6747	1.7025	4.6266
Weighted mean	1.4824	2.1352	1.2964	1.8893	1.8388	3.4850

To assess how realistic our estimates are, we compared them with the typical averages for Europe reported in the meta study conducted by Wardman et al. (2012). The meta-study reports the number of studies considered, the average values, and the corresponding standard errors. Assuming the values used to calculate the reported means follow a normal distribution, then 68%, 95% and 99% of the respective values should be within one, two, and three standard errors (SEs), respectively (Doane and Seward, 2015, Grafarend, 2006). Table 4-5 summarises the estimated range of values for Europe.

A comparison of the GSM and the GPS valuation estimates shows that those of the former are within the expected range of values for Europe, albeit on the upper side probably due to the relatively higher socio-economic status of Switzerland (World Bank, 2018). This is not only another indicator of the relatively high quality of the GSM versus the GPS data used in this study but also gives some reassurance that GSM data can be used for understanding departure time choice. Indeed, these results show that mobile phone data can be feasibly used to analyse time-of-travel choices despite the lack of information on the preferred

departure/arrival times. The availability of demographic data offers additional benefits in terms of explaining the differences in sensitivity across individuals.

Table 4-5 Comparison with time valuations from other sources

Values for Europe (Wardman et al., 2012)			GSM data	GPS data
Description	**Schedule delay early**	**Schedule delay late**		
Mean – 3SE	0.39	1.22		
Mean - 2SE	0.53	1.38		
Mean - SE	0.67	1.54		
Mean	0.81	1.70		
Mean + SE	**0.95**	**1.86**		
Mean + 2SE	**1.09**	**2.02**	1.84 (From Table 4-4)	3.49 (From Table 4-4)
Mean + 3SE	**1.23**	**2.18**		

SE – Standard Error = 0.14 (schedule delay early) and 0.16 (schedule delay late)

4.6 Summary and conclusions

This paper started by analysing the strengths and weaknesses of GPS and GSM data. An initial comparison of the GSM and the GPS datasets collected in parallel for the same users showed that the amount of time gaps in the GPS data were very substantial. This was potentially due to technical issues such as signal losses in urban environments and large public transport vehicles, as well as the users turning off their GPS apps due to battery issues. On the other hand, the amount of time gaps in the GSM data were not as pronounced. Due to these challenges, the GPS data could not capture most of the trips made, and the extracted sample size was very small compared to that extracted from the GSM data. Consequently, the models based on GPS data were not as reliable as those based on GSM data despite the superiority of GPS data. An important point to note is that advances in smartphone GPS technology have occurred since 2010, and it is likely that some of the technical issues encountered in this study have been resolved. Therefore, it is would be important to re-evaluate the feasibility of GPS data using more current datasets. Nevertheless, this paper has successfully demonstrated the potential of GSM data as an alternative source of information for departure time choice modelling.

An important aspect we recognise is the fact that the preferred departure/arrival times of the users are not known due to the anonymous nature of mobile phone data and yet this is an important aspect of departure time choice models. We propose a modelling framework in which the unobserved preferred departure times are assumed to vary across the users following a particular distribution, and estimate the distribution parameters (i.e. the mean and standard deviation) using the mixed logit framework. This approach allows us to simultaneously estimate the distribution parameters alongside the sensitivities to schedule delay, which are found to be intuitive. Although we have applied the proposed approach in the context of anonymous mobile phone data, it can be applied to model departure time choice using traditional RP datasets where the desired times of travel are sometimes not known.

Furthermore, since mobile phone data only reports the revealed departure time preferences of the users, the modeller does not know the other alternatives that were considered while making these choices. We make a general assumption that all the possible departure time intervals in the analysis period (i.e. morning or evening peak period) were considered. However, the attributes for some of the departure time intervals may not be observed for some users if they rarely travel during those periods. Consequently, we propose a practical approach for imputing the travel times associated with different departure time intervals using time-period specific congestion factors derived from Google maps for a sample of O-D pairs. This approach is particularly beneficial in the absence of Google Distance Matrix API services for duration in traffic.

The model results reflect the generally expected behaviour. When these results were applied to estimate the time valuations of schedule delay, we obtained reasonable estimates from the models based on GSM data in comparison with those from the literature, which are largely based on stated choice data. We conclude that the results of this study serve as a proof-of-concept that mobile phone network records are a promising source of information for transport modelling and policy analysis, especially in contexts where traditional data sources are unavailable. This is the case in most developing countries with limited budgets for transport studies.

Acknowledgements

The research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia. We would like to thank the Economic and Social Research Council (ESRC) of the UK and the Institute for Transport Studies, University of Leeds for funding this research. Professor Stephane Hess' time is supported by the European Research Council through the consolidator grant 615596-DECISIONS.

References

Abkowitz, M. D. 1981. An analysis of the commuter departure time decision. *Transportation,* 10**,** 283-297.

Bates, J., Shepherd, N., Roberts, M., Van Der Hoorn, A. & Pol, H. A model of departure time choice in the presence of road pricing surcharges. 18th PTRC Summer Annual Meeting, 1990 University of Sussex, United Kingdom.

Ben-Akiva, M. & Abou-Zeid, M. 2013. Methodological issues in modelling time-of-travel preferences. *Transportmetrica A: Transport Science,* 9**,** 846-859.

Bhat, C. R. 1998a. Accommodating flexible substitution patterns in multi-dimensional choice modeling: formulation and application to travel mode and departure time choice. *Transportation Research Part B: Methodological,* 32**,** 455-466.

Bhat, C. R. 1998b. Analysis of travel mode and departure time choice for urban shopping trips. *Transportation Research Part B: Methodological,* 32**,** 361-371.

Bhat, C. R. 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological,* 35**,** 677-693.

Brey, R. & Walker, J. L. 2011. Latent temporal preferences: An application to airline travel. *Transportation Research Part A: Policy and Practice,* 45**,** 880-895.

Bwambale, A., Choudhury, C. F. & Hess, S. 2017. Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*.

Calastri, C., Hess, S., Choudhury, C., Daly, A. & Gabrielli, L. 2017. Mode choice with latent availability and consideration: theory and a case study. *Transportation Research Part B: Methodological*.

Chen, C., Gong, H., Lawson, C. & Bialostozky, E. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice,* 44**,** 830-840.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Cosslett, S. 1977. The trip-timing decision for travel to work by automobile: demand model estimation and validation. *The Urban Travel Demand Forecasting Project Phase I Final Report, Volume 5.* Institute for Transportation Studies, University of California, Berkeley.

Daly, A. 2010. Cost damping in travel demand models: Report of a study for the

Department for Transport. United Kingdom: RAND Corporation.

Daly, A., Gunn, H., Hungerink, G., Kroes, E. & Mijjer, P. Peak-period proportions in large-scale modelling. 18th PTRC Summer Annual Meeting, 1990 University of Sussex, United Kingdom.

De Groote, A. 2005. GSM Positioning Control. *University of Fribourg, Switzerland***,** 13.

De Jong, G., Daly, A., Pieters, M., Vellay, C., Bradley, M. & Hofman, F. 2003. A model for time of day and mode choice using error components logit. *Transportation Research Part E: Logistics and Transportation Review,* 39**,** 245-268.

Doane, D. P. & Seward, L. E. 2015. *Applied statistics in business and economics,* New York, McGraw-Hill Education.

Doyle, J., Hung, P., Kelly, D., Mcloone, S. F. & Farrell, R. 2011. Utilising mobile phone billing records for travel mode discovery.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. 2011. Hierarchical clustering. *Cluster Analysis, 5th Edition*, 71-110.

Gong, H., Chen, C., Bialostozky, E. & Lawson, C. T. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*, 36, 131-139.

Google Maps. 2018. *Lausanne, Switzerland* [Online]. Google. Available: https://www.google.co.uk/maps/place/Lausanne,+Switzerland/@46.5284586,6.5824556,12z/data=!4m5!3m4!1s0x478c293ecd89a7e5:0xeb173fc9cae2ee5e!8m2!3d46.5196535!4d6.6322734 [Accessed 03 May 2018].

Grafarend, E. W. 2006. *Linear and nonlinear models: fixed effects, random effects, and mixed models*, Walter de Gruyter.

GSM Association. 2017. *The Mobile Economy 2017* [Online]. Available: https://www.gsmaintelligence.com/research/?file=9e927fd6896724e7b26f33f61db5b9d5&download [Accessed 04 November 2017].

Hess, S., Daly, A., Rohr, C. & Hyman, G. 2007a. On the development of time period and mode choice models for use in large scale modelling forecasting systems. *Transportation Research Part A: Policy and Practice*, 41, 802-826.

Hess, S., Polak, J. W. & Bierlaire, M. Functional approximations to alternative-specific constants in time-period choice-modelling. Transportation and Traffic Theory: Flow, Dynamics and Human Interaction, Proceedings of the 16th International Symposium on Transportation and Traffic Theory, 2005. 545-564.

Hess, S., Polak, J. W., Daly, A. & Hyman, G. 2007b. Flexible substitution patterns in models of mode and time of day choice: new evidence from the UK and the Netherlands. *Transportation*, 34, 213-238.

Hydén, C., Nilsson, A. & Risser, R. 1999. How to enhance WALking and CYcliNG instead of shorter car trips and to make these modes safer. Public. Deliverable D6. Walcyng Contract No: UR-96-SC. 099. Department of Traffic Planning and Engineering, University of Lund, Sweden & FACTUM Chaloupka, Praschl & Risser OHG, Vienna, Austria.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40, 63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A. & Willinger, W. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012. Acm, 239-252.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and

opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. & Laurila, J. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*.

Koppelman, F. S., Coldren, G. M. & Parker, R. A. 2008. Schedule delay impacts on air-travel itinerary demand. *Transportation Research Part B: Methodological,* 42**,** 263-273.

Kristoffersson, I. & Engelson, L. 2018. Estimating preferred departure times of road users in a large urban network. *Transportation,* 45**,** 767-787.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J. & Miettinen, M. The mobile data challenge: Big data for mobile computing research. Pervasive Computing, 2012.

Marschak, J. 1960. Binary Choice Constraints on Random Utility Indications. *In:* ARROW, K. (ed.) *Stanford Symposium on Mathematical Methods in the Social Science.* Stanford, California: Stanford University Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics***,** 105-142.

McFadden, D. & Train, K. 2000. Mixed MNL models for discrete response. *Journal of applied Econometrics,* 15**,** 447-470.

Murtagh, F. 1985. Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985*.

NCO. 2018. *Official U.S. government information about the Global Positioning System (GPS) and related topics: GPS Accuracy* [Online]. National Coordination Office for Space-Based Positioning, Navigation, and Timing. Available: https://www.gps.gov/systems/gps/performance/accuracy/ [Accessed 01 June 2018].

Nguyen, D.-Q. 2018. *How work has evolved for Switzerland's women and men* [Online]. swissinfo.ch. Available: https://www.swissinfo.ch/eng/society/gender-roles-since-1970_how-work-has-evolved-for-switzerland-s-men-and-women/43953426 [Accessed 14 June 2018].

Peer, S., Knockaert, J., Koster, P., Tseng, Y.-Y. & Verhoef, E. T. 2013. Door-to-door travel times in RP departure time choice models: An approximation method using GPS data. *Transportation Research Part B: Methodological,* 58**,** 134-150.

Qu, Y., Gong, H. & Wang, P. Transportation mode split with mobile phone data. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 285-289.

Sanko, N., Hess, S., Dumont, J. & Daly, A. 2014. Contrasting imputation with a latent variable approach to dealing with missing income in choice models. *Journal of choice modelling,* 12**,** 47-57.

Schlaich, J. 2010. Analyzing route choice behavior with mobile phone trajectories. *Transportation Research Record: Journal of the Transportation Research Board***,** 78-85.

Schulz, D., Bothe, S. & Körner, C. Human mobility from gsm data-a valid alternative to gps. Mobile data challenge 2012 workshop, June, 2012. 18-19.

Small, K. A. 1982. The scheduling of consumer activities: work trips. *The American Economic Review,* 72**,** 467-479.

Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society***,** 409-424.

Switzerland Tourism. 2018. *Business hours* [Online]. Switzerland Tourism. Available: https://www.myswitzerland.com/en-gb/business-hours.html [Accessed 15 June 2018].

The Economist. 2018. *The glass-ceiling index: Progress has been slow but steady* [Online]. The Economist Group Limited. Available: https://www.economist.com/graphic-detail/2018/02/15/the-glass-ceiling-index [Accessed 15 June 2018].

Themakart 2017. Key Figures. *Urban Audit portraits 2013: core cities.* Neuchâtel, Switzerland: Swiss Federal Statistical Office, ThemaKart.

TomTom. 2016. *Tomtom Traffic Index - Measuring Congestion Worldwide* [Online]. TomTom International BV. Available: https://www.tomtom.com/en_gb/trafficindex/list?citySize=ALL&continent=ALL&country=CH [Accessed 26 May 2018].

Wardman, M., Chintakayala, P., De Jong, G. & Ferrer, D. 2012. European wide meta-analysis of values of travel time. *ITS, University of Leeds, Paper prepared for EIB.*

World Bank. 2018. *GDP per capita (current US $)* [Online]. The World Bank Group. Available: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=PL-GR-PT-DE-EU-CH-GB [Accessed 13 November 2018].

Chapter 5
Getting the best of both worlds - a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling

Andrew Bwambale[*], Charisma F. Choudhury[*], Stephane Hess[*], Md. Shahadat Iqbal[**]

Abstract

Traditional approaches to travel behaviour modelling primarily rely on household travel survey data, which is expensive to collect, resulting in small sample sizes and infrequent updates. Furthermore, such data is prone to reporting errors which can lead to biased parameter estimates and subsequently incorrect predictions. On the other hand, mobile phone call detail records (CDRs), which report the timestamped locations of mobile communication events have been successfully used in the context of generating travel patterns. However, due to their anonymous nature, such records have not been widely used in developing mathematical models establishing the relationship between the observed travel behaviour and influencing factors such as the attributes of the alternatives and the decision makers. In this paper, we propose a joint modelling framework that utilises the advantages offered by both travel survey data and low-cost CDR data to optimise the prediction capacity of traditional trip generation models. In this regard, we develop a model that jointly explains the reported trips for each individual in the household survey data and ensures that the aggregated zonal trip productions are close to those derived from CDR data. This framework is tested using data from Dhaka. Bangladesh consisting of household survey data (65419 persons in 16750 households), mobile phone CDR data (over 600 million records generated by 6.9 million users), and aggregate census data. The model results show that the proposed framework improves the spatial and temporal transferability of the joint models over the base model which relies on household travel survey data alone. This serves as a proof-of-concept that augmenting travel survey data with mobile phone data holds significant promise for the travel behaviour modelling community, not only by saving the cost of data collection, but also improving the prediction capacity of the models.

Keywords: Trip generation, CDR data, mobile phone data, household travel survey data, census data, population synthesis, transferability, Bangladesh, developing country

[*] Choice Modelling Centre, Institute for Transport Studies, University of Leeds (UK)
[**] Lehman Centre for Transport Research, Department of Civil and Environmental Engineering, Florida International University (USA)

5.1 Introduction

Trip generation models form the first component of the traditional four-stage model and are essential for evaluating the relative importance of different factors that influence trip production (Ortúzar and Willumsen, 2011). Being the first step in the four-stage model, the accuracy of trip generation models is critical to that of the subsequent stages.

Traditional approaches to developing trip generation models rely on household travel surveys to establish the mathematical relationship between trip making and the socio-demographics of trip makers (see Bwambale et al., 2015 and the cited references). However, household surveys are often affected by low response rates, small sample sizes and trip reporting errors (e.g. Rolstad et al., 2011, Groves, 2006). Consequently, trip generation models designed to fit household travel survey data alone are likely to result in biased parameters capturing the noise in the data rather than the actual relationships in the population. Aggregating such models to estimate the zonal trip productions often leads to errors, with serious consequences for the subsequent steps of the four-stage model. Furthermore, the high cost of collecting travel survey data makes it difficult makes it difficult to conduct regular model updates.

On the other hand, there has been growing interest in the use of mobile phone data for mobility modelling over the last few decades. Among the various transport related applications, such data has been widely used to estimate origin-destination matrices (e.g. Çolak et al., 2015, Iqbal et al., 2014, Pan et al., 2006, White and Wells, 2002) and trip generation (e.g. Çolak et al., 2015). Since mobile phone data generally covers significant proportions of the population (GSM Association, 2017), the data is able to reliably capture the aggregate travel patterns.

However, due to its anonymous nature, mobile phone data is not traditionally used in developing mathematical models of travel behaviour that establish the relationship between the observed travel behaviour and causal factors such as the attributes of the alternatives and the decision makers. The existing mobility models based on mobile phone data alone cannot be used to reliably test alternative or future travel demand scenarios, and yet this is one of the cardinal roles of transport models. This motivates this research where we combine household travel survey data, aggregate census data, and mobile phone data to jointly optimise the aggregate and the disaggregate fit of trip generation models. In terms of the aggregate fit, we seek to minimise the error between the modelled and the zonal trip productions derived from call detail record (CDR) data, while in terms of the disaggregate fit, we seek to ensure that the model parameters represent the genuine sensitivities of individuals in the population.

Household travel survey data is the most reliable source of information on individual travel behaviour sensitivities (i.e. the model parameter signs and/or the relative magnitudes), however, its limitations can lead to biased parameter scales as mentioned earlier. This prompts us to investigate various ways of adjusting the parameter scales during the joint optimisation process. We adopt a joint optimisation approach because CDR data too is inherently noisy, and thus not error-free. In essence, it would be unrealistic to benchmark one dataset over the other.

The proposed joint modelling framework is both sequential and simultaneous. The approach is sequential in the sense that a base trip generation model is first estimated using household

travel survey data alone to obtain the parameter priors (i.e. the sensitivities). The approach is simultaneous in the sense that when the parameter scales are being adjusted (without changing the prior parameter signs), the model jointly explains the reported trips for each individual in the household survey data and ensures that the aggregated zonal trip productions are close to those derived from CDR data. This ensures that the joint model does not lose the travel behaviour sensitivities reflected in the household survey data.

The rest of the paper is organised as follows, section 5.2 presents a brief review of the literature, section 5.3 presents the data used in this study, section 5.4 presents the modelling framework, section 5.5 presents the model results, and section 5.6 presents the summary and conclusions of the study.

5.2 Literature review

This section presents a brief review of the literature on related work in applying mobile phone data to mobility studies, as well as an overview of different population synthesis techniques.

5.2.1 Related studies on mobile phone data and population synthesis

The availability of large-scale mobile phone data over the last few decades has motivated a lot of research in quantifying human mobility and activity patterns using synthetic data generation methods (e.g. Chen et al., 2014).

From an epidemiology perspective, Vogel et al. (2015) combined CDR data with synthetic populations to model the spread of Ebola in West African countries and obtained promising results with respect to the Ebola predictions of the Centre for Disease Control and Prevention (CDC). Still in West Africa, Cárcamo et al. (2017) developed an intelligent epidemiology simulation software based on synthetic populations comprised of agents with realistic travel behaviour derived from CDR data. In France, Panigutti et al. (2017) compared the spread of a simulated epidemic using CDR and census survey travel patterns, finding greater similarity in areas with high population and connectivity, potentially due to the higher calling rates.

In the field of transport, Zilske and Nagel (2014) generated artificial CDR data from synthetic passengers in a simulated traffic scenario and re-used the data to approximate the amount of missed traffic at different calling rates to quantify the error introduced by CDR location discontinuities. The study found that the errors were inversely proportional to the calling rates and proposed scaling procedures based on observed data such as traffic counts. This led to a subsequent study where simulated CDR data and a synthetic population were combined with link traffic counts to generate all-day trip chains (Zilske and Nagel, 2015). An interesting outcome of this study was that even highly biased CDR data could reasonably reproduce the traffic state across different time periods. The approach of using observed traffic counts to scale CDR data has also been tested in Dhaka in the context of transient origin-destination (OD) matrix estimation (Iqbal et al., 2014).

Still in the field of transport, population synthesis has been applied on real-world mobile phone datasets. Ros and Albertos (2016) developed an improved version of MATSim (an agent-based multi-simulation software) by fusing census and CDR data from Spain to generate synthetic populations with mobility patterns observed in the CDR data. It may be noted that in this particular case, the mobile operator also provided the age and the gender

of the users, which ensured a reliable dependence structure between the travel patterns and socio-demographics in the final synthetic population. However, mobile phone data is usually anonymous, which makes direct socio-demographic linkage impossible. In our earlier work (Bwambale et al., 2017), we developed a demographic group prediction model based on mobile phone usage behaviour extracted from CDR data (as part of a latent class model for trip generation), and can potentially be used for generating synthetic populations, however, this also requires a sub-sample of CDR data with known demographics, which is rarely available.

Kressner (2017) combined consumer and anonymous mobile phone data (wireless signalling and GPS data) from the United States to generate synthetic individual-level trip diaries. The socio-demographics in the disaggregate consumer data were benchmarked against the marginal census totals, while the synthetic travel was benchmarked against the mobility patterns extracted from the aggregate mobile phone data of several operators. A related study was conducted by Zhanga et al. (2017) in the context of social networks in urban simulations. Although these approaches perform quite well in terms of aggregate-level validation, the disaggregate dependency structure in the data seems arbitrary.

To maintain the underlying dependence structure, Janzen et al. (2017) combined household travel survey data, register data (national statistics) and CDR data from France to correct the under-reporting of long-distance trips in travel surveys using population synthesis techniques. The socio-demographics in the travel survey data were matched against those in the register data, while the reported long-distance trips in the travel survey data were matched against those derived from the CDR data. However, a potential issue with this approach is that it assumes uniform under-reporting for all the respondents in the travel survey data, and yet this might vary, at least across different demographic groups, with some cases of over-reporting. Furthermore, the assumed higher reliability of CDR data versus travel survey data is contentious and needs to be approached impartially. This is why we propose an optimisation approach between the two datasets.

5.2.2 Existing methods of population synthesis

Population synthesis is widely applied in activity-based models, and various techniques have been proposed to do this. This section presents a brief review of these methods.

The most widely applied technique is iterative proportional fitting (IPF), which works by fitting a contingency table based on disaggregate survey data to the marginal totals in aggregate census data, constrained by a set of control variables (Beckman et al., 1996). Since its development, various improvements based on the original concept have been proposed to enhance its applicability to new challenges. These improvements have mainly focussed on addressing the zero-cell problem (Guo and Bhat, 2007), simultaneous control of household and individual-level attribute distributions (Casati et al., 2015, Zhu and Ferreira Jr, 2014, Ye et al., 2009, Guo and Bhat, 2007), improving the computational speeds (Pritchard and Miller, 2012), and non-integer conversion to integers (Choupani and Mamdoohi, 2015) etc.

Another popular technique is combinatorial optimisation, which focusses on selecting a subset of households in the disaggregate sample data that closely fit the marginal distributions in the census data for the same area (Voas and Williamson, 2000). This is done by randomly selecting an initial subset of households from the sample data, and iteratively

replacing these with those remaining in the sample data, if and only when this leads to improvements in the fit of the subset. Although this approach has been reported to be superior (Ryan et al., 2009), the IPF method remains the most popular due to its low data requirements, reliability, and faster optimisation (Choupani and Mamdoohi, 2015, Sun and Erath, 2015).

Besides the two methods above, other techniques have been proposed including, the sample-free method (Barthelemy and Toint, 2013), Markov chain Monte Carlo simulation (Farooq et al., 2013), and the Bayesian network framework (Sun and Erath, 2015), among others.

5.3 Data

This section describes the study area, the data used, and the data processing conducted prior to model estimation. The study combines different data types (i.e. household travel survey data, census data, and CDR data) collected at different times between 2009 and 2012. Despite this limitation, these periods are considered close enough to facilitate cross-comparison.

5.3.1 Data description

5.3.1.1 Study area

The study location is Dhaka Metropolitan Area (DMA) in Bangladesh. The area covers approximately 303 square kilometres and is one of the world's most crowded places with a population density of 30551 persons per square kilometre (BBS, 2013). Due to the high population density, the cell tower density is also very high. The area is served by 1361 towers, with most these located in the central business district. The average tower-to-tower distance is approximately 1 kilometre (Iqbal et al., 2014). The total daily trip production from DMA residents was approximately 20.8 million in 2010, with 85.46% of these being home-based (JICA, 2010).

5.3.1.2 CDR data

The CDR data used in this study was provided by Grameenphone Ltd and covers the working days (i.e. Mondays to Thursdays) between 24 June 2012 and 07 July 2012 (2 weeks). The dataset comprises of 6.9 million anonymous users, who together generated over 600 million records during this period (see an excerpt of the CDR data in Table 5-1).

Table 5-1 Excerpt of the CDR data

Unique ID	Date	Time	Duration	Tower Longitude	Tower latitude
AAH03JACKAAAgfBALW	20120624	13:41:49	15	23.9339	90.2931
AAH03JAC8AAAbZfAHB	20120624	13:41:25	73	23.7931	90.2603
AAH03JAC4AAAcvbABC	20120624	13:27:39	8	23.7761	90.4261
AAH03JAC9AAAbWFAVM	20120624	13:27:27	41	23.7097	90.4036
AAH03JABkAAHvEkAQE	20120624	13:32:38	530	23.7386	90.4494

5.3.1.3 Household travel survey data

The household travel survey data used was collected between March 2009 and March 2010 as part of the Dhaka Urban Transport Network Development Study (JICA, 2010). The sampling of households in each zone was based on the population shares at a rate of approximately 1%. The total sample comprises of 65419 individuals and 16750 households, representing an average household size of approximately four persons. The collected information includes each individual's socio-demographic details (e.g. gender, age, working status, income, household size and housing type) and a single day trip diary. Table 5-2 presents the summary statistics of the data.

Table 5-2 Summary statistics of the household survey data

Gender		**Age**		**Working status**		**Trip rate shares**	
Male	53%	0-9 years	15%	Employed	35%	0 trips	43%
Female	47%	10-14 years	9%	Unemployed	38%	1-2 trips	41%
		15-19 years	8%	Student	27%	3-4 trips	14%
		20-29 years	22%			5+ trips	2%
		30-49 years	32%				
		50-59 years	8%				
		60+ years	5%				

5.3.1.4 Census data

The 2011 Bangladesh Population and Housing Census data was used (BBS, 2012). The Census was conducted from 15 to 19 March 2011. The available data reports the aggregate totals of selected person and household level attributes at different geographical scales (e.g. village, ward, and zone (Thana)). Since we could not access the detailed census data due to privacy reasons, we used population synthesis techniques (Ye et al., 2009) to generate realistic artificial populations for the different study area zones by combining the aggregate census data with the household survey data as explained later in Section 5.3.2.2. It may be noted that the fusion of household survey data and census data could only be done at the zone (Thana) level due to differences in the study area delimitations at smaller geographical scales. In total, we successfully matched 31 zones. The variables available in both datasets are summarised in Table 5-3.

Table 5-3 Variables in both the census and the household survey data

Data	**Household survey data**	**Census data**
Individual attributes	Gender	Population by gender
	Age-group	Population by age-group
	Working status *(employed, unemployed, student)*	Population by working status
	Occupation *(agriculture, industry, services)*	Population by occupation
Household attributes	Household size	Number of households by household size
	Household type *(permanent, semi-permanent, thatched etc.)*	Number of households by household type

5.3.2 Data processing and combination

5.3.2.1 General concept

The overarching idea is to minimise the difference between the zonal trip productions derived from CDR data and those obtained by aggregating the disaggregate trip generation model, without compromising the behavioural sensitivities reflected in the household survey data. Model aggregation is based on a synthetic population generated using the Iterative Proportional Updating technique (Ye et al., 2009). Figure 5-1 presents a summary of the data processing framework. The subsequent sections discuss the key aspects of this framework.

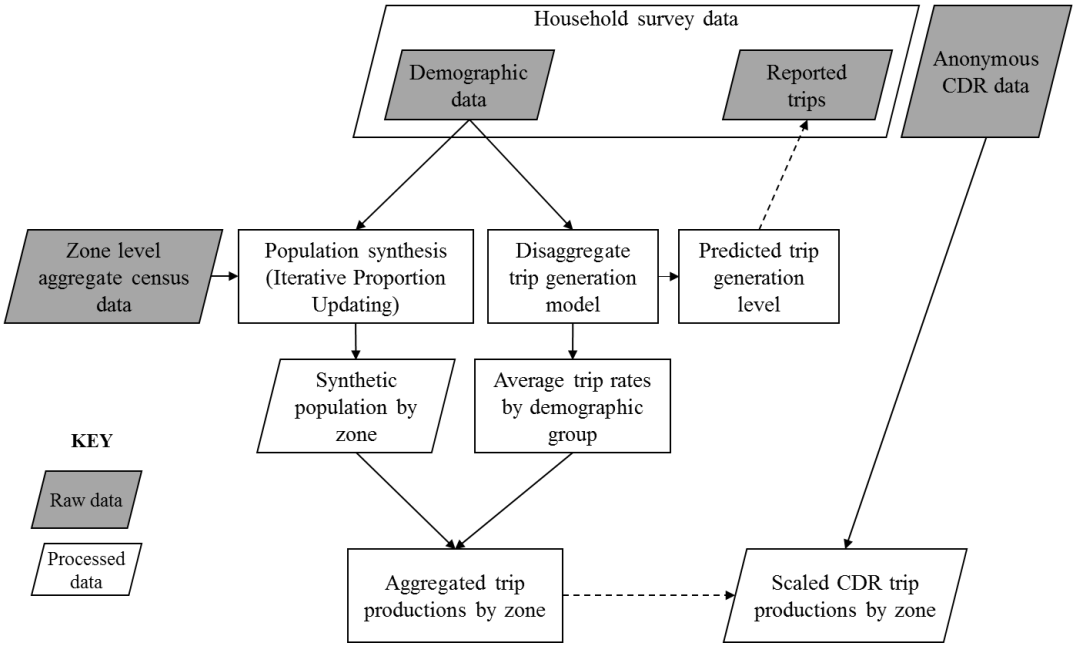


Figure 5-1 Data processing framework

5.3.2.2 Population synthesis

Among the various software applications for population synthesis, we used PopGen (Ye et al., 2009), which is capable of conducting Iterative Proportional Updating (IPU). This algorithm simultaneously controls for both the person and the household-level attribute distributions during the fitting procedure, and has been proven to perform better than the simpler methods.

The algorithm relies on person and the household-level sample data for each zone (derived from the household survey data) to generate zone-specific synthetic populations, which are validated against the corresponding person and the household-level aggregate census totals.

Tables 5-4 and 5-5 present lists of the control variables used in PopGen. It may be noted that we did not use the individual's occupation as we could not reliably match the categories in the household survey and the census data.

Table 5-4 Household-level control variables used in PopGen

HSETYP	**Housing type**	**HHLDSIZE**	**Household size**
HSETYP1	Pucka (Permanent house)	HHLDSIZE1	1
HSETYP2	Semi-pucka (Semi-permanent house)	HHLDSIZE2	2
HSETYP3	Kutcha (Thatched house)	HHLDSIZE3	3
HSETYP4	Jhupri (Slum house)	HHLDSIZE4	4
		HHLDSIZE5	5
		HHLDSIZE6	6
		HHLDSIZE7	7
		HHLDSIZE8	8+

Table 5-5 Individual-level control variables used in PopGen

GEND	**Gender**	**AGEP**	**Age-group**
GEND1	Male	AGEP1	0-9 years
GEND2	Female	AGEP2	10-14 years
		AGEP3	15-19 years
WRKST	**Working status**	AGEP4	20-29 years
WRKST1	Employed	AGEP5	30-49 years
WRKST2	Unemployed	AGEP6	50-59 years
WRKST3	Student	AGEP7	60+ years

Figure 5-2 presents the distribution of the Average Absolute Relative Differences (AARD) across the zones. This metric gives the mean deviation of the person weighted sums with respect to the household and person aggregate census totals. As observed, the AARD values for most zones are concentrated in the lower ranges of the axis, an indication that the population synthesis was successful.

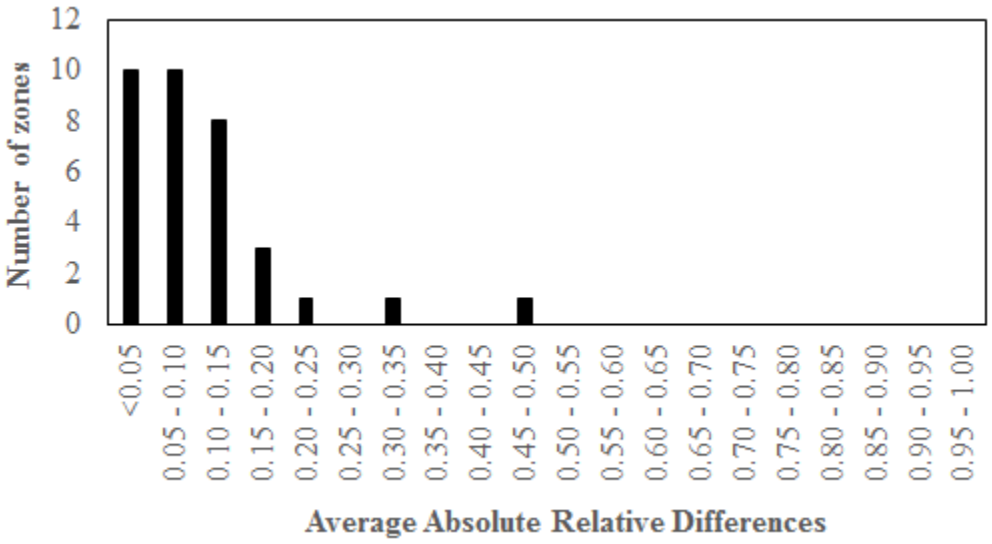


Figure 5-2 Distribution of the AARD values

Furthermore, comparisons of the synthetic versus the actual estimates for each attribute at the person and the household levels are presented in Figures 5-3 and 5-4 respectively, where the distributions are observed to closely match.

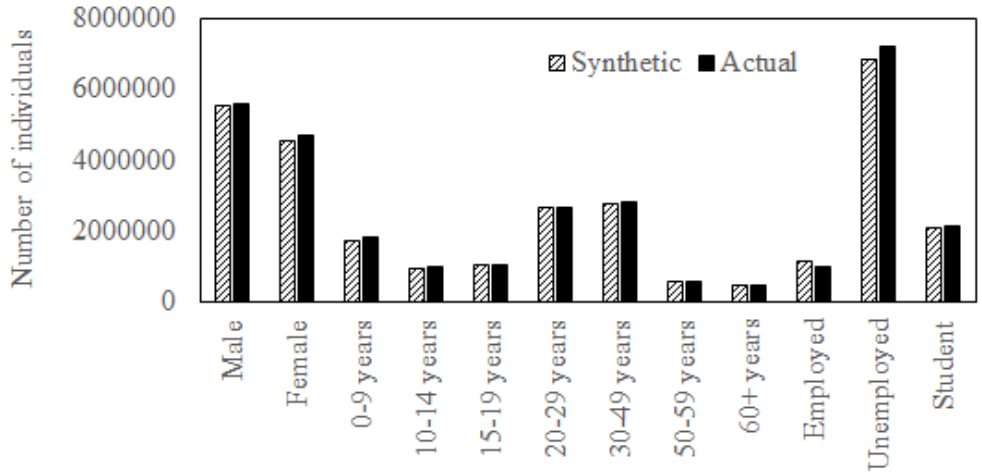


Figure 5-3 Distribution of the individual-level estimates

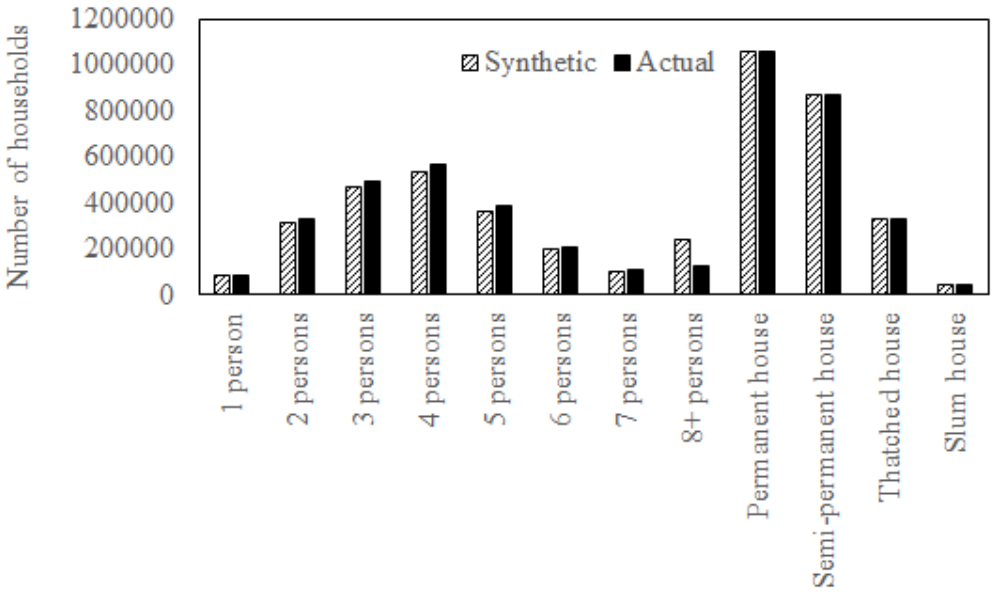


Figure 5-4 Distribution of the household-level estimates

5.3.2.3 Extraction of zonal trip productions from CDR data

The CDR data for the entire observation period was first analysed to identify each user's home location, which was defined as the most frequently observed cell tower at night (i.e. between 8 pm and 6 am). The labelled cell towers (i.e. home/others) for each user were then arranged according to the date and observation timestamp.

Home-based trips were extracted by considering any two consecutive CDR events from different cell towers, with one of those being the home cell tower. After conducting several trials, a lower distance threshold of 0.5 kilometres between subsequent towers was considered as the optimum for minimising the number of false trips due to tower jumps.

An upper threshold of 24 hours or midnight (whichever came first) was specified based on the assumption that a user typically travels from and back to home within the same effective day. Consequently, if the first and the last CDR events for the day were not at the home cell tower, corresponding raw trips were added (Çolak et al., 2015).

The number of users and home-based trips associated with each cell tower were recorded. The cell towers within the boundaries of each zone were then grouped with the aid of GIS software (QGIS Development Team, 2018). The total trips for each zone were then corrected using the ratio of the zonal population to the number of users classified as residents of the zone (Çolak et al., 2015).

5.4 Modelling framework

We propose an approach that combines two modelling strategies, that is, discrete choice modelling at the individual level and ordinary least squares at the aggregate level.

5.4.1 Individual-level trip generation model (Base model)

Discrete choice models have been the most preferred approach for modelling trip generation over the last few decades (e.g. Bwambale et al., 2015, Pettersson and Schmöcker, 2010, Agyemang-Duah and Hall, 1997). Although the ordered response choice mechanism has been the most preferred approach for modelling trip generation, previous findings in the context of car ownership choices (which are also ordered) have shown that the unordered response choice mechanism outperforms the former (Bhat and Pulugurta, 1998). To implement the unordered response choice mechanism, we rely on the random utility theory (Marschak, 1960). Let U_{nt} be the utility of individual n making t trips. This can be expressed as;

$$U_{nt} = \beta_t' X_n + \varepsilon_{nt} \tag{5-1}$$

Where X_n is a vector of the socio-demographic attributes of individual n, β_t is a vector of the model parameters to be estimated, and ε_{nt} is the random component of utility. Since the individual socio-demographics are constant across the alternatives, we specify a different set of parameters for each trip generation level to reflect the fact that each attribute has a differential impact on the utility for each trip generation level.

Under the assumption that the error terms (ε_{nt}) are distributed independently and identically across alternatives and individuals using a type I extreme value distribution, the trip generation choice probabilities can be calculated using the multinomial logit (MNL) model (McFadden, 1974) as expressed below;

$$P_{nt} = \frac{\exp(\beta_t' X_n)}{\sum_{t^*} \exp(\beta_{t^*}' X_n)} \tag{5-2}$$

Where P_{nt} is the probability of individual n making t trips.

If we were to rely on the household travel survey data alone, the model parameters would be estimated by maximising the log-likelihood function below.

$$LL(\beta_t) = \sum_n \sum_t K_{nt} \ln(P_{nt}) \tag{5-3}$$

Where dummy variable $K_{nt} = 1$ if and only if individual n makes t trips, otherwise $K_{nt} = 0$.

However as mentioned earlier, fitting the model to match the trips reported in the household travel survey data alone can lead to biased parameter estimates due to reporting errors, thereby resulting in misrepresentation of the aggregate travel demand as reflected in Figure 5-5, where the predicted aggregate zonal trips from the base model are different from those derived from the CDR data, especially towards the right hand side of the figure. The proposed joint modelling framework (in the next section) seeks to optimise such differences.

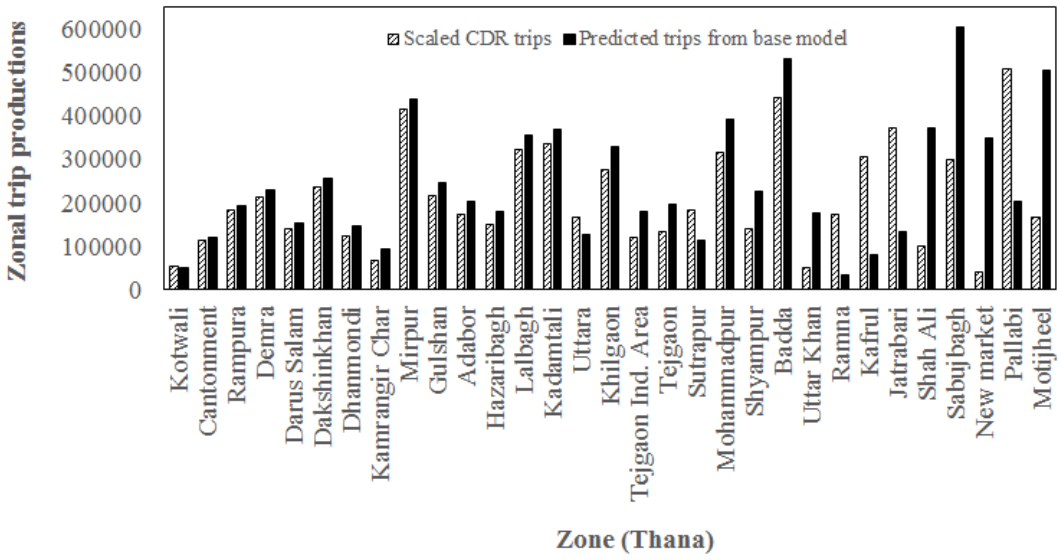


Figure 5-5 Distribution of the CDR trip productions

5.4.2 Joint trip generation model

The framework of the joint trip generation model is both sequential and simultaneous. It is sequential as it relies on the pre-estimated base model to obtain the priors of the parameter signs and relative magnitudes. However, when the parameter scales are being adjusted (without changing the prior parameter signs), the joint model simultaneously optimises performance at both the aggregate and disaggregate levels with respect to the CDR and the household travel survey data, respectively.

As mentioned earlier, this combined approach ensures that the resulting model does not lose the travel behaviour sensitivities reflected in the household travel survey data, by maintaining the sensitivities from the base model. Adjusting the parameter scales has an impact on the choice probabilities for each trip generation outcome, which influences the expected trip rates of the individuals. The framework of the joint trip generation model is described below. Let $\widehat{U}_{nt}$ be the updated utility of individual n making t trips. This can be expressed as;

$$\widehat{U}_{nt} = \alpha\beta_t' X_n + \varepsilon_{nt} \qquad (5\text{-}4)$$

Where α is a vector of the scaling factors to be estimated. The β parameters are priors derived from the base model, and are not re-estimated in the joint framework. The specification of the scaling factors is discussed later on.

The updated trip generation choice probability can be expressed as follows;

$$\hat{P}_{nt} = \frac{\exp(\alpha\beta'_t X_n)}{\sum_{t^*}\exp(\alpha\beta'_{t^*} X_n)} \quad (5\text{-}5)$$

Where $\hat{P}_{nt}$ is the updated probability of making t trips by individual n.

However, to estimate the scaling factors, we need to fulfil two objectives. The first objective is to explain the reported trips for each individual in the household survey data. The second objective is to ensure that the aggregated zonal trip productions are close to those derived from CDR data. Both outcomes have a probability attached to them and the simultaneous estimation maximises the joint probability of the two outcomes.

To estimate the aggregate zonal trip productions, we rely on the synthetic population generated in section 5.3.2.2. As mentioned earlier, the synthetic population was designed to match both the person and the household-level attribute distributions during the fitting procedure, thus making it more reliable. We have a synthetic population of M simulated individuals identified as m with $m = 1, \ldots, M$, and a study area comprising of Z zones identified as z with $z = 1, \ldots, Z$. Let $\hat{P}_{mt}$ denote the updated probability of making t trips by simulated individual m. It may be noted that $\hat{P}_{mt}$ is equivalent to $\hat{P}_{nt}$ if both the simulated individual and the actual respondent in the household survey data have the same demographics (i.e. the values of $\hat{P}_{mt}$ depend on the calculations of $\hat{P}_{nt}$). Now, let $\hat{T}_z$ denote the aggregate zonal trip production for zone z. This can be calculated by taking the weighted average trips for each simulated individual, in which the updated MNL probabilities are the weights, and summing across the zonal synthetic population as follows;

$$\hat{T}_z = \sum_{m=1}^{M}\left[Y_{mz}\left(\sum_{t=1}^{T}(t * \hat{P}_{mt})\right)\right] \quad (5\text{-}6)$$

Where dummy variable $Y_{mz} = 1$ if and only if simulated individual m belongs to zone z, otherwise, $Y_{mz} = 0$. The objective is to ensure that $\hat{T}_z$ is as close as possible to the corrected CDR trip productions for zone z. If φ_z denotes the corrected CDR trip productions for zone z, the relationship between φ_z and $\hat{T}_z$ can be expressed as follows;

$$\varphi_z = \hat{T}_z + \omega_z \quad (5\text{-}7)$$

Where ω_z is an error term which we assume follows a normal distribution with a mean of zero, $\omega_z \sim N(0, \sigma^2)$. $P(\varphi_z)$ is then the likelihood of observing the CDR trip productions for zone z, and, from Equation 5-7, this can be expressed as follows;

$$P(\varphi_z) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\varphi_z - \hat{T}_z)^2}{2\sigma^2}\right) \quad (5\text{-}8)$$

$P(\varphi_z)$ clearly depends on $\hat{P}_{nt}$ given that $\hat{T}_z$ is a function of $\hat{P}_{mt}$, which depends on the calculations of $\hat{P}_{nt}$ as explained earlier. For each survey respondent in zone z, we need to maximise the probability of the chosen alternative and ensure that the probabilities of all the alternatives maximise $P(\varphi_z)$. Let t_n^o denote the number of trips observed for individual n in the household survey data, such that $\hat{P}_{nt^o}$ gives the logit probability of the observed

choice for individual n. The overall joint likelihood (L) of the observed choices and the aggregate CDR trip productions across individuals is calculated as follows;

$$L = \prod_{n=1}^{N}\left[\sum_{z=1}^{Z} H_{nz}\left(\hat{P}_{nt^o} * P(\varphi_z)\right)\right] \tag{5-9}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{n=1}^{N}\left[\sum_{z=1}^{Z} H_{nz}\left(\frac{\exp\left(\alpha\beta'_{t^o}X_n\right)}{\sum_{t^*}\exp\left(\alpha\beta'_{t^*}X_n\right)} * \exp\left(\frac{-(\varphi_z - \hat{T}_z)^2}{2\sigma^2}\right)\right)\right]$$

Where dummy variable $H_{nz} = 1$ if and only if survey respondent n belongs to zone z.

Since products are difficult to differentiate, we obtain the log-likelihood (LL) by applying logarithms to Equation 5-9 resulting in Equation 5-10.

$$LL = -\frac{N}{2}log(2\pi) - Nlog(\sigma) + \tag{5-10}$$

$$\sum_{n=1}^{N}\sum_{z=1}^{Z} H_{nz}\left(\ln\left[\frac{\exp\left(\alpha\beta'_{t^o}X_n\right)}{\sum_{t^*}\exp\left(\alpha\beta'_{t^*}X_n\right)}\right] - \frac{1}{2\sigma^2}(\varphi_z - \hat{T}_z)^2\right)$$

Three parameter scaling scenarios are tested, and these are;

- Model 1

  This specification applies the same α scaling factor to the utility models of the different trip generation levels (see Equation 5-4), i.e. $\alpha_t = \alpha, \forall t$. The updated utility models have the same relative variable sensitivities as in the base model, albeit with different parameter scales.

- Model 2

  This specification applies a different α_t scaling factor to the utility model of each trip generation level. The updated utility models maintain the base model relative variable sensitivities for each particular trip generation level, however, the variable sensitivities across the different trip generation levels are adjusted with different parameter scales, and hence the relative values across levels change from the base model.

- Model 3

  This specification applies a different α_x scaling factor to each explanatory variable X (e.g. gender, age-group, and working status), however, α_x does not change across the different trip generation levels. The updated utility models maintain the base model attribute-level relative sensitivities for a particular variable across the different trip generation levels, however, the inter-variable relative sensitivities are adjusted with different parameter scales.

5.4.3 Model evaluation framework

The performance of the joint models is evaluated in terms of both the temporal and the spatial transferability as illustrated in Figures 5-6 and 5-7, respectively.

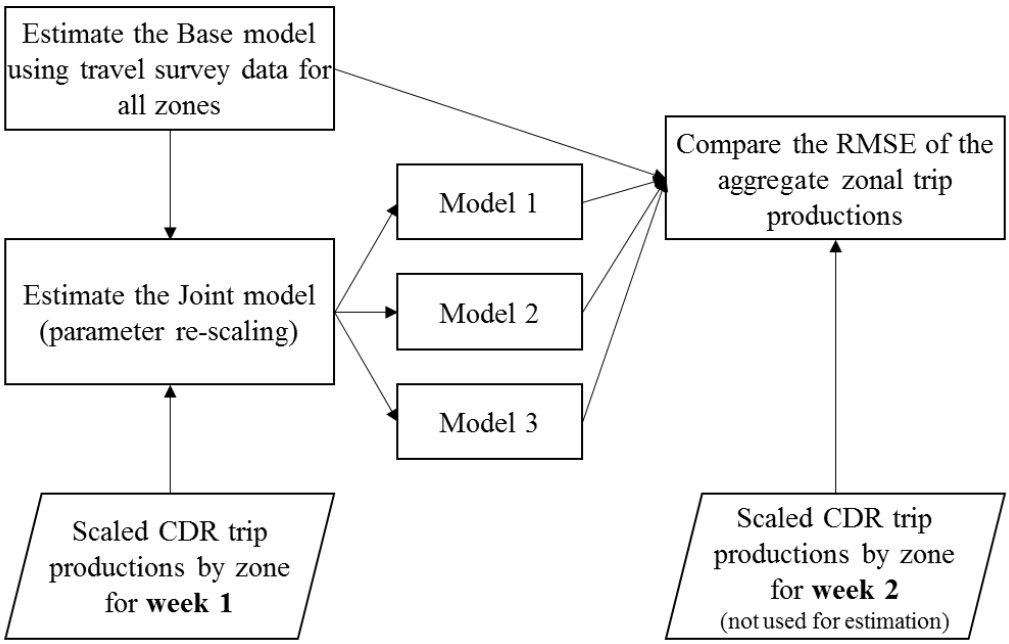


Figure 5-6 Temporal transferability framework

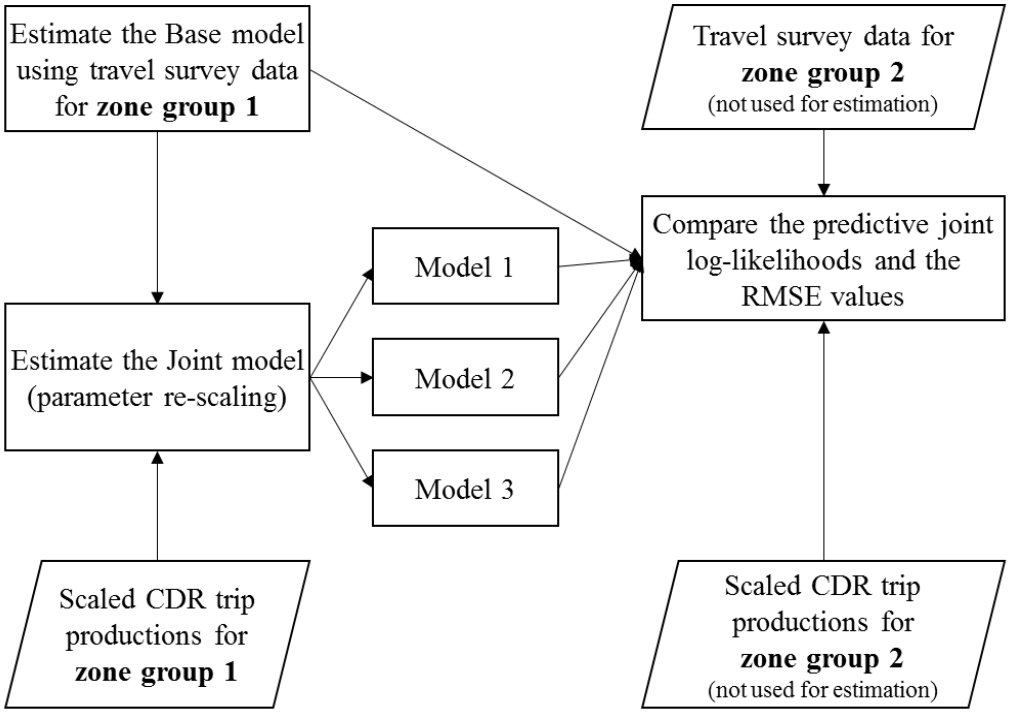


Figure 5-7 Spatial transferability framework

In terms of temporal transferability, the joint models associated with each parameter scaling scenario are estimated using the zonal aggregate CDR trip productions for week 1. The prediction capacities of the estimated joint models, as well as the base model are then compared in terms of the root mean square errors with respect to the zonal aggregate CDR trip productions for week 2 (see Figure 5-6).

In terms of spatial transferability, the study area zones are randomly divided into two groups. The base and the joint models are then estimated using the data for one group of zones and applied to the other group of zones (not used for estimation). The prediction capacities of the models are then compared in terms of the predictive joint log-likelihoods, and the root mean square errors with respect to the aggregate CDR trip productions of the application zones (see Figure 5-7).

5.5 Modelling results

This section presents the final model specification, as well as the model estimation and validation results.

5.5.1 Variable specification

The dependent variable is the number of individual home-based trips (irrespective of the trip purpose). This is because we could not reliably infer the purposes of the CDR trips. Based on distributions in the data, the trip generation levels were grouped into 0, 1-2, 3-4, and 5+ trips per day.

The explanatory variables considered for possible inclusion in the model are those that were used for population synthesis. The household-level variables (i.e. household size and type) were however not included in the final model as they led to unreasonable parameter signs, potentially due to their weak influence on individual trip-making decisions. The final model specification thus contains the gender, the age-group, and the working status of the individuals, coded as dummy variables.

For model identification purposes, the parameters associated with the zero trip generation level were treated as the base (for all explanatory variables). Furthermore, male non-workers in the 30-49 age-group were treated as the base demographic group, and their preferences are entirely explained by the alternative specific constants. Thus, the model parameter estimates represent the differential impact on utility with respect to the zero trip generation level and the base demographic group.

5.5.2 Estimation results

5.5.2.1 Base model

We first estimated the base model to assess whether the parameter estimates are in line with the expected travel behaviour. The model results are presented in Table 5-6.

The alternative specific constants capture the underlying differential impact on utility with respect to the zero trip generation level. All the estimates are negative, and their magnitude increases with respect to the trip generation level. Keeping all other factors constant, this reflects a general tendency to make fewer trips, especially by the base category (i.e. male, non-workers, aged 30-49 years).

Table 5-6 Base model results

Variable	Parameter	t-statistic
Alternative specific constants (ASCs)		
1-2 trips	-0.2069	-7.46
3-4 trips	-1.0408	-24.56
5+ trips	-3.0859	-31.19
Dummies specific to gender (base category is males)		
Females		
1-2 trips	0.0870	3.94
3-4 trips	-0.2841	-7.95
5+ trips	-0.2654	-3.15
Dummies specific to working-status (base category is non-workers)		
Workers		
1-2 trips	0.4630	17.23
3-4 trips	0.9252	23.05
5+ trips	1.1482	12.38
Students		
1-2 trips	1.4079	46.47
3-4 trips	0.9381	17.13
5+ trips	-0.5333	-2.65
Dummies specific to age-group (base category is the 30-49 years age-group)		
Age 1-9 years		
1-2 trips	-1.6354	-50.69
3-4 trips	-3.1065	-36.73
5+ trips	-3.5549	-9.46
Age 10-14 years		
1-2 trips	-0.8143	-19.49
3-4 trips	-1.7635	-22.52
5+ trips	-1.9201	-6.00
Age 15-19 years		
1-2 trips	-0.6539	-16.22
3-4 trips	-0.9669	-15.71
5+ trips	-1.0077	-5.71
Age 20-29 years		
1-2 trips	-0.1457	-5.67
3-4 trips	-0.3249	-9.58
5+ trips	-0.3009	-4.02
Age 50-59 years		
1-2 trips	-0.1423	-4.12
3-4 trips	-0.2552	-5.92
5+ trips	-0.3721	-3.81

Table 5-6 cont'd

Variable	Parameter	t-statistic
Age 60+ years		
1-2 trips	-0.2494	-5.63
3-4 trips	-0.3531	-6.14
5+ trips	-0.4853	-3.47
Measures of fit		
Number of observations	65419	
Log-likelihood at zero	-90689.99	
Log-likelihood at convergence	-64859.90	
Number of parameters	30	
Adjusted rho-square	0.2845	
Likelihood ratio	51660.10	
P value of the likelihood ratio	0.0000	

The parameter estimates for females represent the differential impact on utility with respect to males. For 1-2 trips, we obtain a positive parameter estimate, while for the higher trip generation levels, we obtain negative parameter estimates. The proportion of women working in the garments industry, one of the leading sectors in Dhaka, is 64-90% (ADB and ILO, 2016). This probably explains the positive parameter sign for 1-2 trips. Otherwise, males are more likely to make a higher number of trips compared to females, probably due to the average higher income levels of the former (BBS, 2012) and socio-cultural factors.

The parameter estimates for the working status variables (i.e. workers and students) represent the differential impact on utility with respect to non-workers. As observed, the parameters for workers are positive, and their magnitudes increase with respect to the trip generation level, an indication that workers generally make more trips compared to non-workers. On the other hand, the parameter estimates for students are positive for 1-2 and 3-4 trips, and negative for 5+ trips. This shows that students make more trips compared to non-workers only up to a reasonable level expected for school going individuals.

Similarly, the parameter estimates for the age-group variables represent the differential impact on utility with respect to the 30-49 years age-group. As observed, the parameter estimates for all the other age-groups are negative, an indication that they generally make fewer trips compared to the base age-group (30-49 years). The active working age of white-collar workers in Bangladesh typically ranges between 29 and 60 years (i.e. the latest age for completing tertiary education and the retirement age respectively (BBS, 2012)). It is therefore reasonable that persons in the 30-49 years age-group are more active travellers due to their economic vibrancy.

Finally, it is observed that the overall model (in terms of the likelihood ratio), as well as all the parameter estimates (in terms of the t-statistics) are statistically significant at the 99% level of confidence (see Ben-Akiva and Lerman, 1985 for details).

5.5.2.2 Joint models

As mentioned earlier, the parameters of the base model were fixed in the joint modelling framework, and only the scaling factors were estimated. Table 5-7 presents the estimated scaling factors and the measures of fit for all the three models for comparison purposes.

Positive scaling factors were obtained for all the three models, an indication that the resulting coefficients in the scaled joint models have the same signs as those in the base model.

Table 5-7 Joint model scaling factors

Description of scaling factor	Model 1		Model 2		Model 3	
	Estimate	**t-stat**	**Estimate**	**t-stat**	**Estimate**	**t-stat**
Model 1 Uniform factor (applied to all the base model parameters)	1.3650	2280.16				
Model 2 **(Factors specific to trip generation level)**						
1-2 trips			1.2716	131.39		
3-4 trips			1.4873	247.83		
5+ trips			1.1699	158.63		
Model 3 **(Factors specific to particular variables)**						
Gender					1.5228	33.81
Working status					1.8148	105.16
Age-group					1.3262	120.70
ASCs					1.6023	171.51
Measures of fit						
Convergence LL at the disaggregate level	-66002.75		-65914.01		-67747.10	
Convergence LL at the aggregate level	-718560.40		-718377.10		-715805.30	
Joint convergence LL	-784563.20		-784291.20		-783552.40	
Base model convergence LL	-64859.90		-64859.90		-64859.90	
Base model LL at the aggregate level	-805093.10		-805093.10		-805093.10	
Base model joint convergence LL	-869953.00		-869953.00		-869953.00	
Likelihood ratio (joint model w.r.t the base model)	170780		171234		172801	
P value	0.0000		0.0000		0.0000	

A comparison of the joint convergence log-likelihoods shows that Model 3 gives the best performance, followed by Model 2, and then Model 1. This is attributed to the flexibility of the parameter scaling framework. An important point to note is that all the three joint models perform better than the base model in terms of the joint log-likelihood.

As mentioned earlier, during model optimisation, we are basically dealing with a trade-off between disaggregate and aggregate model performance. Thus, the disaggregate log-likelihood of the joint models is a little worse than that of the base model. However, if the base model parameters are directly used to estimate the joint log-likelihood, it is observed that the model yields the worst performance.

The p-values of the likelihood ratios of the joint models with respect to the base model are all less than 0.01, an indication that the improvements in performance are statistically significant at the 99% confidence level beyond the advantages offered by the additional parameters (see Ben-Akiva and Lerman, 1985 for details).

5.5.3 Model evaluation in terms of transferability

The models based on the full sample have been presented in the previous section. To evaluate the stability and the predictive performance of the joint models as well as the base model, we compared their temporal and spatial transferability following the evaluation framework described in Section 5.4.3. Tables 5-8 and 5-9 present the measures of fit in terms of the temporal and the spatial transferability, respectively.

From Table 5-8, it is observed that the temporal transferability of the joint models is generally higher than that of the base model in terms of the joint log-likelihoods and the root mean square errors (RMSE) with respect to the zonal CDR trips. Among the three joint models, Model 3 offers the best transferability, however, Model 2 gives the best prediction at the disaggregate level in both the estimation and the application contexts.

Table 5-8 Temporal transferability

	Measure	Base model	Model 1	Model 2	Model 3
Week 1 (Estimation)	LL (disaggregate level)	-64859.90	-66024.40	-65940.80	-67850.40
	LL (aggregate level)	-805642.50	-719566.80	-719396.20	-716695.30
	Joint LL	-870502.40	-785591.20	-785337.00	-784545.70
Week 2 (Application)	LL (disaggregate level)	-64859.90	-66024.40	-65940.80	-67850.40
	LL(aggregate level)	-804545.50	-717793.90	-717596.20	-715031.60
	Joint LL	-869405.40	-783818.30	-783537.00	-782882.00
	RMSE w.r.t CDR trips	43342.84	13547.09	13527.84	13328.49

For spatial transferability, we tested both directions of model transfer. It may be noted that the general interpretation of the base model parameters for each group of zones did not change. From Table 5-9, it is again observed that the joint models are generally more transferrable compared to the base model in terms of the joint log-likelihoods and the root mean square errors for both directions.

In this particular case, it is observed that Model 2 gave the best disaggregate prediction for the zone group 1 to 2 transfer direction, while Model 1 gave the best disaggregate prediction for the reverse transfer direction.

An important point worth mentioning is that the superior performance of the base model at the disaggregate level is expected as it was designed to fit the travel survey data alone, but as mentioned earlier, this could be prone to reporting errors and hence less dependable.

Table 5-9 Spatial transferability

	Measure	Base model	Model 1	Model 2	Model 3
Zone group 1 (Estimation)	LL (disaggregate level)	-26102.10	-26712.45	-26652.76	-27724.63
	LL(aggregate level)	-321381.60	-290869.40	-290725.20	-288898.10
	Joint LL	-347483.70	-317581.85	-317377.96	-316622.73
Zone group 2 (Application)	LL (disaggregate level)	-38859.38	-39701.58	-39352.09	-41303.51
	LL(aggregate level)	-491580.30	-429017.00	-428604.80	-426638.20
	Joint LL	-530439.68	-468718.58	-467956.89	-467941.71
	RMSE w.r.t CDR trips	50626.73	13375.06	13274.68	13161.58
Zone group 2 (Estimation)	LL (disaggregate level)	-38688.76	-39227.43	-39333.92	-40185.59
	LL(aggregate level)	-482400.40	-428113.30	-427818.70	-426238.10
	Joint LL	-521089.16	-467340.73	-467152.62	-466423.69
Zone group 1 (Application)	LL (disaggregate level)	-26219.53	-26689.06	-26786.11	-27445.95
	LL(aggregate level)	-315772.10	-289862.10	-289890.20	-288799.10
	Joint LL	-341991.63	-316551.16	-316676.31	-316245.05
	RMSE w.r.t CDR trips	38776.13	13702.57	13758.49	13602.58

From the results, it is clear that Model 3 gives the best overall spatial and temporal transferability, however, the disaggregate performance of Models 1 and 2 as highlighted above shows that these parameter scaling approaches offer some benefits as well. These results present initial efforts to exploit the benefits of both household travel survey and mobile phone data to optimise the performance of travel behaviour models, and there is a need for further research using data from different contexts to investigate the different parameter scaling approaches in further detail.

5.6 Summary and conclusions

This paper started by highlighting the reporting errors and sampling bias associated with household travel survey data, and how these could lead to biased model parameters (e.g. Rolstad et al., 2011, Groves, 2006). The paper outlines the possible consequences of such issues in the context of trip generation, where the estimated models would misrepresent the distribution of the aggregate travel demand across zones.

The paper demonstrates the feasibility of a joint modelling framework to find the best fit at both the aggregate and disaggregate levels by combining household travel survey, census, and CDR data. The joint modelling framework operates by adjusting the parameter scale(s) of a pre-estimated base model to jointly optimise the prediction accuracy with respect to the reported trips in travel survey data and the zonal aggregate trip productions derived from CDR data.

Three different approaches of parameter scaling were investigated (i.e. uniform, alternative specific, and variable specific scaling corresponding to joint models 1, 2, and 3 respectively). All the three joint models were found to have higher temporal and spatial transferability compared to the base model which relies on household travel survey data alone, thus making them more reliable. Although variable specific scaling (Model 3) produced the best overall results, there is a need for further research using data from different contexts to investigate if this finding is universally applicable.

Although the proposed framework has been tested in the context of trip generation, it has potential benefits in improving the modelling of other transport choices (such as mode choice, route choice, departure time choice etc.). We conclude that the results of this study serve as a proof-of-concept that mobile phone data can be fused with traditional data sources to improve the temporal and spatial transferability of models. This approach is particularly important in the context of developing countries where reliable traditional data sources are scarce, and models making use of low-cost passive data to enhance their temporal and spatial transferability are invaluable.

Acknowledgements

The research in this paper used mobile phone data made available by Grameenphone Ltd, Bangladesh, household travel survey data provided by the Japan International Cooperation Agency (JICA), and aggregate census data obtained from the Bangladesh Bureau of Statistics (BBS). We would like to thank the Economic and Social Research Council (ESRC) of the UK and the Institute for Transport Studies, University of Leeds for funding this research. Professor Stephane Hess' time is supported by the European Research Council through the consolidator grant 615596-DECISIONS.

References

ADB & ILO 2016. Bangladesh: Looking beyond garments: Employment diagnostic study. Manila, Phillipines: Asian Development Bank and International Labour Organization.

Agyemang-Duah, K. & Hall, F. L. 1997. Spatial transferability of an ordered response model of trip generation. *Transportation Research Part A: Policy and Practice,* 31**,** 389-402.

Barthelemy, J. & Toint, P. L. 2013. Synthetic population generation without a sample. *Transportation Science,* 47**,** 266-279.

BBS 2012. Community Report: Dhaka Zila: June 2012. *Population and Housing Census 2011*. Dhaka: Bangladesh Bureau of Statistics (BBS).

BBS 2013. District Statistics 2011 Dhaka. Dhaka: Bangladesh Bureau of Statistics.

Beckman, R. J., Baggerly, K. A. & Mckay, M. D. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30**,** 415-429.

Ben-Akiva, M. E. & Lerman, S. R. 1985. *Discrete choice analysis: theory and application to travel demand*, MIT press.

Bhat, C. R. & Pulugurta, V. 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological*, 32**,** 61-75.

Bwambale, A., Choudhury, C. F. & Hess, S. 2017. Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*.

Bwambale, A., Choudhury, C. F. & Sanko, N. Modelling Car Trip Generation in the Developing World: The Tale of Two Cities. Transportation Research Board 94th Annual Meeting, 2015.

Cárcamo, J. G., Vogel, R. G., Terwilliger, A. M., Leidig, J. P. & Wolffe, G. Generative models for synthetic populations. Proceedings of the Summer Simulation Multi-Conference, 2017. Society for Computer Simulation International, 7.

Casati, D., Müller, K., Fourie, P. J., Erath, A. & Axhausen, K. W. 2015. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record: Journal of the Transportation Research Board***,** 107-116.

Chen, C., Bian, L. & Ma, J. 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46**,** 326-337.

Choupani, A.-A. & Mamdoohi, A. R. 2015. Population Synthesis in Activity-Based Models: Tabular Rounding in Iterative Proportional Fitting. *Transportation Research Record: Journal of the Transportation Research Board***,** 1-10.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Farooq, B., Bierlaire, M., Hurtubia, R. & Flötteröd, G. 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58**,** 243-263.

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly***,** 646-675.

GSM Association. 2017. *The Mobile Economy 2017* [Online]. Available: https://www.gsmaintelligence.com/research/?file=9e927fd6896724e7b26f33f61db5b9d5&download [Accessed 04 November 2017].

Guo, J. & Bhat, C. 2007. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board***,** 92-101.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies,* 40**,** 63-74.

Janzen, M., Müller, K. & Axhausen, K. W. Population Synthesis for Long-Distance Travel De-mand Simulations using Mobile Phone Data. 6th Symposium of the European Association for Research in Transportation (hEART 2017), 2017.

JICA 2010. Dhaka Urban Transport Network Development Study (DHUTS) in Bangladesh, Final Report. Dhaka: Japan International Cooperation Agency.

Kressner, J. D. 2017. Synthetic Household Travel Data Using Consumer and Mobile Phone Data. *Final Report for NCHRP IDEA Project 184.* Transportation Research Board.

Marschak, J. 1960. Binary Choice Constraints on Random Utility Indications. *In:* ARROW, K. (ed.) *Stanford Symposium on Mathematical Methods in the Social Science.* Stanford, California: Stanford University Press.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics***,** 105-142.

Ortúzar, J. D. D. & Willumsen, L. G. 2011. *Modelling transport,* John Wiley & Sons.

Pan, C., Lu, J., Di, S. & Ran, B. 2006. Cellular-based data-extracting method for trip distribution. *Transportation Research Record: Journal of the Transportation Research Board***,** 33-39.

Panigutti, C., Tizzoni, M., Bajardi, P., Smoreda, Z. & Colizza, V. 2017. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. *Royal Society open science,* 4**,** 160950.

Pettersson, P. & Schmöcker, J.-D. 2010. Active ageing in developing countries?–trip generation and tour complexity of older people in Metro Manila. *Journal of Transport Geography,* 18**,** 613-623.

Pritchard, D. R. & Miller, E. J. 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation,* 39**,** 685-704.

QGIS Development Team. 2018. *QGIS Geographic Information System* [Online]. Available: https://qgis.org/en/site/ [Accessed 14 August 2018].

Rolstad, S., Adler, J. & Rydén, A. 2011. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health,* 14**,** 1101-1108.

Ros, O. G. C. & Albertos, P. G. 2016. D5.4 Enhanced Version of MATSim: Synthetic Population Module. *Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development (INSIGHT).* Madrid, Spain: INSIGHT Consortium.

Ryan, J., Maoh, H. & Kanaroglou, P. 2009. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis,* 41**,** 181-203.

Sun, L. & Erath, A. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies,* 61**,** 49-62.

Voas, D. & Williamson, P. 2000. An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography,* 6**,** 349-366.

Vogel, N., Theisen, C., Leidig, J. P., Scripps, J., Graham, D. H. & Wolffe, G. 2015. Mining Mobile Datasets to Enable the Fine-Grained Stochastic Simulation of Ebola Diffusion. *Procedia Computer Science,* 51**,** 765-774.

White, J. & Wells, I. Extracting Origin Destination Information from Mobile Phone Data. Eleventh International Conference on Road Transport Information and Control (Conf. Publ. No. 486), March 2002 London. IET, pp. 30 - 34.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. 88th Annual Meeting of the Transportation Research Board, Washington, DC, 2009.

Zhanga, D., Caob, J., Feygina, S., Tangc, D. & Pozdnoukhova, A. 2017. Connected Population Synthesis for Urban Simulation. *Personal Communication. Draft Available from Authors by Request*.

Zhu, Y. & Ferreira Jr, J. 2014. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record,* 2429**,** 168-177.

Zilske, M. & Nagel, K. 2014. Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science,* 32**,** 802-807.

Zilske, M. & Nagel, K. 2015. A simulation-based approach for constructing all-day travel chains from mobile phone data. *Procedia Computer Science,* 52**,** 468-475.

Chapter 6
Discussion and conclusions

6.1 Summary

This thesis has advanced a number of methodological and applied contributions around the theme of improving the behavioural and policy underpinnings of transport models based on mobile phone network records. The aim was to extend the application of mobile phone data beyond the visualisation of anonymous mobility patterns by incorporating the data into traditional transport modelling approaches to quantify the relative importance of the underlying factors influencing the observed travel behaviour. The thesis offers remedies to some of the shortcomings that have thus far prevented such applications by fusing the data with information from external sources. The research findings show that the data has great potential as it is more representative, large, and frequent, thereby capturing more variability over long periods of time, which opens up a lot of scope for validation as demonstrated in some sections of this thesis.

The introduction section highlighted the gaps in the literature and outlined the research objectives. This section discusses the progress made in achieving these objectives across the different chapters of the thesis, the contributions to knowledge and practice, the potential beneficiaries of the current research findings, and the future research directions. The section ends by providing some concluding remarks.

6.2 Progress made in achieving the research objectives

This section revisits each of the research objectives highlighted in the introduction section. The progress made is discussed by highlighting the linkages between the general and the specific objectives realised in each of the four chapters as presented earlier in Table 1-8, reproduced below as Table 6-1 for easy cross-referencing.

Table 6-1 Linkages between the overall research goals and the specific objectives (Copy of Table 1-8)

	Specific objectives and the corresponding chapters			
General objectives	**S1** **Chapter 2** *Trip generation*	**S2** **Chapter 3** *Route choice*	**S3** **Chapter 4** *Departure time choice*	**S4** **Chapter 5** *Trip generation*
G1	✓	✓	✓	✓
G2	✓			✓
G3		✓	✓	
G4	✓	✓	✓	✓

G1 Developing innovative methods for combining mobile phone network records with traditional data sources to facilitate the analysis of travel behaviour

Data fusion was a common aspect across the work done in this thesis. The models in chapters 3 and 4 were estimated using the attributes of the alternatives (i.e. the routes in chapter 3, and the departure times in chapter 4), while those in chapters 2 and 5 were estimated using demographic variables.

Although imputing the attributes of the alternatives is a challenging process, once they are determined, it is not difficult to assign them to the corresponding alternatives. Chapter 3 focussed on modelling long-distance route choice, in which one of the key explanatory variables is travel cost. This was computed in terms of the vehicle operating costs (fuel and non-fuel costs) using the HDM-III model (Watanatada et al., 1987), an earlier version of the more advanced HDM-4 model (Kerali, 2000), which we could not use due to input data constraints. Given the anonymous nature of the data, the vehicle operating costs per user were computed as weighted averages using information on the typical vehicle occupancy rates and the mode shares of the study area. In chapter 4, where the main focus was departure time choice modelling, imputing the travel times for the unobserved time intervals was the main challenge, and these were estimated with the aid of the Google maps direction tool as outlined in the previous section.

On the other hand, incorporating demographic variables is more challenging, and was done through a system of probabilistic relationships. Both chapters 2 and 5 focus on this issue, though following different approaches. The hybrid modelling framework developed in chapter 2 combines the demographic details, the CDR data, and the GSM data for a sub-sample of users to develop a latent class trip generation model. In a different way, the joint trip generation modelling framework developed in chapter 5 combines household travel survey data, aggregate census data, and CDR data to re-adjust the parameter scales of a disaggregate trip generation model to optimise both its aggregate and disaggregate reliability. Here, the parameters in the disaggregate trip generation model are probabilistically linked to the zonal CDR trip productions through a synthetic population generated for the study area using iterative proportional updating (Ye et al., 2009).

From all the four cases above, reasonable results were obtained, an indication that data fusion is a feasible way of understanding the travel behaviour characteristics of a study area despite the anonymous nature of the data.

G2 Evaluating the shortcomings of traditional modelling approaches and proposing mitigation measures using mobile phone network records to optimise the reliability and applicability of the models

This objective was met in chapters 2 and 5 of the thesis. The work in Chapter 2 outlines the estimation and application challenges associated with traditional modelling approaches in the context of trip generation. In terms of estimation, the chapter briefly highlights the limitations associated with traditional travel survey data with regard to reporting errors, while in terms of application, it underscores the constraints related to the lack of detailed demographic data in CDR data to feed into the models to predict aggregate travel demand.

In response to these challenges, a novel hybrid modelling framework was developed. The framework comprises of two sequential sub-models. The first sub-model is a demographic prediction model, which is estimated by calibrating the mobile phone usage behaviour of a sub-sample of the users against their reported demographics. Information on mobile phone usage is extracted from the user's CDR data. The demographic group membership probabilities from the first sub-model are then used as class weights inside a latent class model for trip generation (the second sub-model). The trip generation model is calibrated using trip rates extracted from GSM data, which is more appropriate for trip generation compared to CDR data.

The framework takes advantage of the fact that GSM data captures all the trips made by users with active mobile phones, except the short trips within the GSM cell boundaries. This implies that the data is more suitable for urban areas where the GSM cell sizes are small (for example less than 1 km), thereby leading to fewer missed trips. The use of GSM data mitigates the need for trip diaries.

The framework recognises that demographic and GSM data are usually not available on a large scale (due to privacy concerns and very large storage requirements respectively), and avoids such limitations during model application as it only depends on anonymous CDR data. Since CDR data is usually stored by mobile network operators for billing purposes, the proposed framework provides a low-cost alternative to predict trip rates on a large scale.

The work in chapter 5 focuses on optimising both the aggregate and disaggregate reliability of traditional trip generation models using a novel joint modelling framework. The chapter draws motivation from previous studies that have highlighted the challenges of travel survey data collection in terms of the sampling bias, survey response burden, and reporting errors (e.g. Rolstad et al., 2011, Groves, 2006). These limitations imply that disaggregate models based on travel survey data alone are likely to produce unreliable parameter estimates. As a result, such models may not satisfactorily represent the aggregate travel demand patterns of the study area. Thus, low-cost CDR data, which can give a fair representation of the aggregate travel demand patterns, is used to optimise the reliability of the models.

However, summing the disaggregate model outcomes requires detailed demographic data of the study area population, which is never available due to privacy reasons. This prompts the use of population synthesis techniques to generate artificial populations matching the demographic distributions in both the household survey and the aggregate census totals for the different zones in the study area (Ma, 2011, Kirill and Axhausen, 2011, Ye et al., 2009, Pritchard, 2008). The disaggregate trip generation model is applied to the zonal synthetic populations to estimate the zonal trip productions, which are then compared against those extracted from the CDR data.

The proposed joint modelling framework recognises that CDR data too is not error-free, and optimises model performance at both the aggregate and disaggregate levels by updating the parameter scales without changing the behavioural dynamics reflected in the household survey data. The framework has been tested on data for Dhaka, and results show that it improves both the temporal and spatial transferability of the models, thus making them more reliable.

G3 Analysing the limitations of mobile phone network records with regard to specific modelling scenarios and developing appropriate methods to deal with those limitations

Discussions about the general limitations of mobile phone data and how to address them are a common theme across all the chapters of this thesis. However, this objective was specifically achieved in chapters 3 and 4. In chapter 3, CDR data was used to analyse long-distance route choice behaviour. Being event-driven, the data reports discontinuous mobile phone locations, which makes it impossible to observe the full trajectories of the users. Instead, only the partial trajectories can be observed, and this depends on the mobile phone usage rate during a particular trip. For very close OD pairs, there is even an increased possibility of not capturing the partial trajectories as users may travel from the origin to the destination without using their phones. Thus, the chapter argues that CDR data is more suitable for long-distance trips, where there is an increased likelihood of phone usage during the journeys. The chapter further notes that with the increasing usage of mobile internet data services, which is also reflected in CDR data, the associated location discontinuities are likely to reduce in the near future, thus making the data suitable for short trips as well.

The limitation of only observing the partial trajectories poses challenges for route choice behaviour analysis in a highly overlapping network. A route assignment algorithm that labels the extracted partial trajectories as either unique or shared across a group of routes was developed. The labelled trajectories were then used to analyse route choice behaviour by adapting the broad choice framework, which was developed in the context of vehicle type choice (Wong, 2015), to the current modelling scenario.

In this particular case, the attributes of the alternatives, rather than those of the users, were used as explanatory variables due to the anonymous nature of the data. These were imputed from various sources as explained in the chapter. Different models to account for the overlapping nature of the network were tested (i.e. the c-logit and the path-size logit models). A comparison of these models against the base MNL model showed that CDR data is able to capture the expected behaviour towards overlapping routes, with the path-size logit model giving the best performance. The parameter estimates for the path-size logit model were reliable as they produced realistic estimates of the value of travel time for the study area.

However, the route assignment algorithm may have limitations in dense inter-urban networks, where it would be difficult to observe a small enough subset of the possible routes using few CDR locations. Nevertheless, this limitation is likely to be overcome in the near future with the increasing trend of mobile internet usage (Gerpott and Thomas, 2014), which will increase the frequency of the CDR locations.

Chapter 4 presents another related scenario. The chapter started by critically analysing the strengths and weaknesses of GPS versus GSM data, and how these impact the model outcomes (in the context of departure time choice modelling). An interesting finding was that the GSM data was more reliable than the GPS data despite the superior location accuracy of the latter. This was mainly caused by the big time gaps in the GPS data, potentially due to technical reasons such as signal losses in urban environments and large public transport vehicles, as well as the users turning off their GPS apps due to battery depletion (NCO, 2018, Gong et al., 2012, Chen et al., 2010). Notwithstanding the possible

influence of poor smartphone GPS technology in 2009/2010 (i.e. the data collection period), the chapter highlights the need to always conduct quality checks on different types of big data prior to adopting any of them.

Besides analysing the strengths and weaknesses of the different datasets, both were used for modelling departure time decisions, which presented two challenges. Firstly, the desired times-of-travel of the users were unknown, secondly, the travel times could only be observed for the chosen departure time intervals. For the first challenge, an assumption was made that the desired times-of-travel vary randomly across the users, and a mixed logit framework (see Train, 2009 for details) was developed to estimate the mean and the standard deviation of the distribution. For the second challenge, a practical approach was developed to impute the unobserved travel times using time-period specific congestion factors estimated with the aid of the Google Maps direction tool, which predicts the average travel times between a given O-D pair at different departure or arrival times (Google Maps, 2018).

From the model estimation, it was found that the time gaps in the data had an impact on the reliability of the results. This was reflected in the valuation metrics derived from both models, where those obtained from the GSM data were closer to those based on traditional data sources. The developed approaches can be used in different modelling scenarios, other than those discussed in this thesis.

G4 Assessing the potential of models based on mobile phone network records to capture the expected travel behaviour in terms of the parameter estimates and/or policy insights in terms of the derived valuation metrics

The discussion in this section is presented in two parts starting with the capacity to capture reasonable travel behaviour in terms of the parameter estimates followed by the potential to obtain realistic policy insights in terms of the derived valuation metrics.

G4-1 Travel behaviour in terms of the parameter estimates

The work in chapters 2, 3, 4, and 5 involves the estimation of travel behaviour models. In each of these chapters, the model parameters were discussed in terms of whether they are in agreement with the expected travel behaviour.

Beginning with chapter 2, the parameters of the traditional trip generation model based on the observed demographics and trip rates extracted from GSM data produced intuitive parameter signs. The parameters in the traditional trip generation model were used as a reference for assessing those in the hybrid trip generation model, and it was found that the latter produced similar parameter signs as the former. Furthermore, the differences in the parameter magnitudes were not statistically significant. This shows that the hybrid framework is able to capture the same travel behaviour as it would be in a traditional model, thereby leading to similar policy conclusions.

Likewise, the parameters of the long-distance route choice models in chapter 3 were assessed for compliance with expected travel behaviour, and it was found that they were all in agreement with a priori expectations. However, more interesting were the parameter signs of the systematic utility correction factors aimed at accounting for overlap (i.e. the commonality and the path size terms). In both cases, the expected parameter signs were

obtained, an indication that CDR data is able to capture the expected behaviour towards overlapping routes.

In chapter 4, the departure time choice models based on both the GSM and the GPS data had the expected parameter signs. For the model based on GSM data, the differences in the travel time and the schedule delay sensitivities by gender and age-group were evaluated and found to be reasonable for the study area.

Finally, in chapter 5, the main focus was to re-scale the parameters of a base model estimated using household travel survey data. In this case, the base model parameters were fixed, and only the scaling factors were estimated. In order to maintain the behavioural sensitivities reflected in the household survey data, the estimated scaling factors would need to be positive. The fact that all the estimates are positive shows that the distribution of the zonal trip productions obtained from the base model is not sufficiently different from that extracted from the CDR data to cause a change in the overall travel behaviour.

In summary, the findings from all the four modelling scenarios show that mobile phone data is able to capture the expected travel behaviour sensitivities. These findings demonstrate the potential of mobile phone data as an alternative data source for developing transport models.

G4-2 Policy insights in terms of the derived valuation metrics

This objective was achieved in chapters 3 and 4, where the model results were used to estimate the value of travel time and the time valuation of schedule delay respectively. Both metrics are useful in transport policy appraisal.

In chapter 3, the estimated values of travel time were found to be close to the median wage of the study area. Although the median wage is not necessarily equivalent to the value of travel time, it gives a good indication of the range in which these values should fall, thus the obtained results are promising. Similarly, in chapter 4, the estimated time valuations of schedule delay based on the GSM data (which was more reliable compared to GPS data) were found to be close to those sourced from similar studies based on traditional data.

These findings serve as a proof-of-concept that mobile phone data can be used for policy analysis, especially in contexts where traditional data sources are not available.

6.3 Contributions to knowledge and practice

This research has extended the application of mobile phone data to travel behaviour modelling. The progress made in achieving the research objectives was discussed in section 6.2. This section summarises the key contributions of the research to the field of transport.

6.3.1 Extending the application of mobile phone network records to travel behaviour modelling and policy analysis

In general, the use of mobile network records for developing econometric models of travel behaviour is still very low. At the moment, the only study found to do this is by Schlaich (2010). This thesis provides major advances along this research direction and is the first to use CDR and GSM data to estimate valuation metrics, which are important in transport policy appraisal. The findings in this thesis motivate further research into exploring the hidden potential of big data to solve real-world transport policy problems.

6.3.2 A hybrid modelling framework to address the issue of unobserved user demographics in transport models based on mobile phone data

This framework was developed in chapter 2 in the context of trip generation modelling. The benefits of the framework in terms of addressing the estimation and application challenges associated with traditional modelling approaches on big data with missing socio-demographic information have been outlined. The proposed framework can be applied to mitigate similar problems in different fields of transport modelling (such as route choice, departure time etc.), as well as beyond transport, for example in health and general consumer choice modelling using big data.

6.3.3 A novel joint modelling framework for optimising the aggregate and disaggregate performance of models

This framework was developed in chapter 5 in the context of trip generation modelling. The motivation was to mitigate the effects of sampling bias and reporting errors in travel survey data, which can lead to biased parameter estimates (e.g. Rolstad et al., 2011, Groves, 2006). The proposed framework demonstrates the potential of mobile phone data to solve this common problem, and can generally be applied to analogous multi-objective optimisation problems in different modelling scenarios.

6.3.4 Applying the broad choice framework to the context of route choice modelling using noisy CDR data

The long-distance route choice model developed in chapter 3 applies the broad choice modelling framework, which was developed in the context of vehicle type choice (Wong, 2015). This is the first application of such a framework to a route choice modelling scenario. The framework was used to leverage the limitations of CDR data where unique route choices could not be observed for some users, and only the broad sub-groups of the possible routes were identifiable. This demonstrates the opportunities that are available to adapt established approaches from other fields to solve problems in analysing big data. Numerous other applications are possible, including for example, route choice in a public transport network where only the entry and exit points of users are observed, such as with many smartcard systems.

6.3.5 A new method for modelling departure time choice without information on the desired times-of-travel

Information on the desired times-of-travel is critical for modelling departure time choice decisions, however, such data is usually not available in anonymous records. The weaknesses of previously developed approaches to address this problem were highlighted in chapter 4, and a new approach was proposed to overcome such limitations. The proposed approach is unique in the sense that it allows the modeller to understand the sensitivities, as well as the valuations attached to schedule delay, despite the passive nature of the data. The approach can be applied to traditional revealed preference datasets, where the preferred departure times are sometimes also not reported.

6.4 Potential beneficiaries of the research findings

The application of big data to transportation studies remains one of the defining research challenges in this era. Therefore, the obvious beneficiaries of this study are researchers interested in exploring the hidden potential of emerging big data sources. The developed methods are likely to inspire further innovation around the theme of travel behaviour modelling using big data sources.

Furthermore, government transport agencies around the world are increasingly becoming aware of the emerging opportunities presented by big data to deliver efficient and smart transport solutions. For example, the UK government is currently supporting research related to big data adoption for transportation studies through its Universities and the recently established Transport Systems Catapult (Hill et al., 2017, Transport Systems Catapult, 2015, Hanley and Hobbs, 2014). The Ministry of Transport in Argentina is currently involved in research engagements with the World Bank to develop tools for collecting and analysing big data for better transport planning, and have so far developed a system that generates origin-destination matrices from smart card data (Quiros and Arias, 2018). The Jakarta Provincial Government in partnership with the United Nations Global Pulse have recently completed a study on real-time data analytics using bus GPS data to improve the efficiency of public transport operations (UN Global Pulse, 2017). Similarly, Dalberg Data Insights in partnership with Kampala Capital City Authority have developed a transport mobility application for travel pattern analysis using mobile phone data (Dalberg Data Insights, 2018). With the growing interest in big data across both the developed and the developing worlds, the innovative data fusion and modelling frameworks developed in this research are likely to be of great importance to government agencies in terms of analysing travel behaviour and policy formulation.

Although emerging big data sources have generated interest in both the developed and the developing worlds, the latter context is expected to experience relatively higher benefits as traditional data sources are commonly unavailable due to limited budgets for data collection. At the moment, most investment decisions are driven by approximations as opposed to models calibrated with real-world data. Transport authorities in such contexts are likely to look more towards these emerging data sources for better investment decisions. Thus the findings of this research are particularly timely for transport practitioners in developing countries, which are now being referred to as the global south.

Finally, it is worth noting that this research is also likely to have impacts beyond transport. For example, the hybrid and the joint modelling frameworks presented in chapters 2 and 5 respectively can be applied in the field of health to identify high-risk groups and model the spread of epidemics by fusing mobile phone data, census data, and demographic and health survey data.

6.5 Future research directions

This section briefly discusses insights into possible directions of future research. The first obvious step is to apply the methods developed in this thesis to different contexts using the latest available mobile phone datasets. In particular, investigating the performance of these methods in more densely populated urban areas with complex transport networks and travel patterns would be a good contribution. Furthermore, most applications tested in this thesis

did not have local models for comparison purposes. Therefore, campaigns to collect primary travel survey data alongside mobile phone data for validation purposes are necessary.

The next logical step is to compare the predictions based on the developed models against those based on machine learning techniques. Machine learning techniques have already been applied in related transport studies (e.g. Ellis et al., 2014, Wang et al., 2010, Farrahi and Gatica-Perez, 2008, Sohn et al., 2006), however, as earlier mentioned, they are mainly suited to prediction, and do not explain the underlying behavioural interrelationships.

Furthermore, in the context of route choice modelling, it would be worth exploring the potential of map-matching techniques (Quddus et al., 2007). These algorithms typically rely on GPS data, whose location accuracy is higher than that of mobile phone network records. Some progress has already been registered in the use of CDR data to develop these algorithms (Han et al., 2018, Algizawy et al., 2017, Schulze et al., 2015). As the spatial and temporal resolution of mobile phone network records improves (for example due to technological advancement and increased mobile internet data usage), it would be of interest to evaluate how this impacts the performance of the different map-matching algorithms.

Research into the development of dynamic econometric models of travel behaviour is another interesting direction for future work. Previous studies have already used mobile phone data to analyse the spatial and temporal distributions of human mobility (e.g. Yuan and Raubal, 2012, Loibl and Peters-Anders, 2012, Calabrese et al., 2011, Saravanan et al., 2011). Incorporating policy-sensitive variables into such models to explain the observed spatial-temporal behaviour would help in the better prediction of network-wide impacts.

Another key direction for future research is choice set determination, particularly in the field of route choice modelling, where several alternative routes can be possible, and yet individuals do not consider all while making choices (Prato, 2009). Observing a user's mobile phone mobility patterns over a long period of time can provide insights into their usual set of possible routes. However, research needs to be done to formalise the process of determining the optimum observation period and the amount of variety seeking behaviour that should be considered.

Finally, additional research into the field of mode detection would be of great interest to policymakers since mode choice is one of the key factors influencing traffic congestion in most cities. The few available studies have yielded promising results (Qu et al., 2015, Doyle et al., 2011, Wang et al., 2010, Reddy et al., 2008, Sohn et al., 2006), however, there is a need for further research to improve the transferability of these approaches to real-world urban scenarios, where they are needed most.

6.6 Concluding remarks

This thesis focussed on using mobile phone network records to develop econometric models of travel behaviour. Four different modelling scenarios were discussed and tested in detail, with all producing results that are in agreement with the expected travel behaviour.

These findings come at a time when mobile phone penetration rates are growing in both the developed and the developing world (GSM Association, 2017). Hence there is a likelihood that future datasets will cover even much wider proportions of the population. Among the various datasets, CDR data is the most readily available, however, it presents the greatest challenge in terms of extracting relevant information for travel behaviour analysis.

Nevertheless, future CDR datasets are likely to be more favourable due to the increasing trend in mobile internet data usage (Gerpott and Thomas, 2014). This will increase the frequency of the captured CDR locations, thus making the data less discontinuous, including for shorter journeys. It is therefore expected that the scope of using mobile phone data as well as the robustness of the developed models will improve with time. Crucially, mobile phone data will allow for more regular updating of models given the continuous data collection, something that will become ever more crucial given the rapid changes to travel patterns.

The thesis demonstrates the potential of mobile phone network records as a low-cost alternative source of information for transport modelling and policy analysis. The methodological and applied contributions made in the research however have the potential of being applied to different modelling scenarios in both the developed and the developing worlds, and beyond transport in general.

Despite the progress made so far, there remains a need for continued research to improve the current approaches and test new ideas. Sustained research collaborations with mobile network operators and other key stakeholders will be crucial in obtaining the data required for further research.

References

Algizawy, E., Ogawa, T. & El-Mahdy, A. 2017. Real-Time Large-Scale Map Matching Using Mobile Phone Data. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 11**,** 52.

Calabrese, F., Di Lorenzo, G., Liu, L. & Ratti, C. 2011. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. *IEEE Pervasive Computing,* 10**,** 36-44.

Chen, C., Gong, H., Lawson, C. & Bialostozky, E. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice,* 44**,** 830-840.

Dalberg Data Insights. 2018. *Dalberg Data Insights Maps Urban Mobility across Haiti and around Kampala, Uganda* [Online]. Dalberg Data Insights. Available: https://www.dalberg.com/our-ideas/dalberg-data-insights-maps-urban-mobility-across-haiti-and-around-kampala-uganda [Accessed 13 November 2018].

Doyle, J., Hung, P., Kelly, D., Mcloone, S. F. & Farrell, R. 2011. Utilising mobile phone billing records for travel mode discovery.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J. & Kerr, J. 2014. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Frontiers in public health,* 2**,** 36.

Farrahi, K. & Gatica-Perez, D. Daily routine classification from mobile phone data. International Workshop on Machine Learning for Multimodal Interaction, 2008. Springer, 173-184.

Gerpott, T. J. & Thomas, S. 2014. Empirical research on mobile Internet usage: A meta-analysis of the literature. *Telecommunications Policy,* 38**,** 291-310.

Gong, H., Chen, C., Bialostozky, E. & Lawson, C. T. 2012. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems,* 36**,** 131-139.

Google Maps. 2018. *Lausanne, Switzerland* [Online]. Google. Available: https://www.google.co.uk/maps/place/Lausanne,+Switzerland/@46.5284586,6.5824556,12z/data=!4m5!3m4!1s0x478c293ecd89a7e5:0xeb173fc9cae2ee5e!8m2!3d46.5196535!4d6.6322734 [Accessed 03 May 2018].

Groves, R. M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly***,** 646-675.

GSM Association. 2017. *The Mobile Economy 2017* [Online]. Available: https://www.gsmaintelligence.com/research/?file=9e927fd6896724e7b26f33f61db5b9d5&download [Accessed 04 November 2017].

Han, B., Tang, X., Hu, Z. & Yu, K. A Map Matching Algorithm for Complex Road Conditions Based on Base Station Data. Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on, 2018. IEEE, 426-431.

Hanley, S. & Hobbs, A. 2014. Big and Open Data in Transport. London: Houses of Parliament - Parliamentary Office of Science and Technology (POST).

Hill, N., Gibson, G., Guidorzi, E., Amaral, S., Parlikad, A. K. & Jin, Y. 2017. Scoping Study into Deriving Transport Benefits from Big Data and the Internet of Things in Smart Cities. *Final Report for Department for Transport.* Oxford: Ricardo Energy & Environment.

Kerali, H. G. R. 2000. *Overview of HDM-4*, Paris, The World Road Association (PIARC), Paris and The World Bank, Washington, DC.

Kirill, M. & Axhausen, K. W. Population synthesis for microsimulation: State of the art. Transportation Research Board 90th Annual Meeting, 2011.

Loibl, W. & Peters-Anders, J. 2012. Mobile phone data as source to discover spatial activity and motion patterns. *G1_Forum***,** 524-533.

Ma, L. 2011. *Generating disaggregate population characteristics for input to travel-demand models.* Doctor of Philosophy, University of Florida.

NCO. 2018. *Official U.S. government information about the Global Positioning System (GPS) and related topics: GPS Accuracy* [Online]. National Coordination Office for Space-Based Positioning, Navigation, and Timing. Available:

https://www.gps.gov/systems/gps/performance/accuracy/ [Accessed 01 June 2018].

Prato, C. G. 2009. Route choice modeling: past, present and future research directions. *Journal of choice modelling,* 2**,** 65-100.

Pritchard, D. R. 2008. *Synthesizing agents and relationships for land use/transportation modelling.* Masters of Applied Science, University of Toronto.

Qu, Y., Gong, H. & Wang, P. Transportation mode split with mobile phone data. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 285-289.

Quddus, M. A., Ochieng, W. Y. & Noland, R. B. 2007. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation research part c: Emerging technologies,* 15**,** 312-328.

Quiros, T. P. & Arias, S. E. 2018. *Data analytics for transport planning: five lessons from the field* [Online]. The World Bank Group. Available: http://blogs.worldbank.org/transport/data-analytics-transport-planning-five-lessons-field [Accessed 13 November 2018].

Reddy, S., Burke, J., Estrin, D., Hansen, M. & Srivastava, M. Determining transportation mode on mobile phones. Wearable Computers, 2008. ISWC 2008. 12th IEEE International Symposium on, 2008. IEEE, 25-28.

Rolstad, S., Adler, J. & Rydén, A. 2011. Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health,* 14**,** 1101-1108.

Saravanan, M., Pravinth, S. V. & Holla, P. Route detection and mobility based clustering. Internet Multimedia Systems Architecture and Application (IMSAA), IEEE 5th International Conference, 2011. IEEE, 1-7.

Schlaich, J. 2010. Analyzing route choice behavior with mobile phone trajectories. *Transportation Research Record: Journal of the Transportation Research Board***,** 78-85.

Schulze, G., Horn, C. & Kern, R. Map-matching cell phone trajectories of low spatial and temporal accuracy. Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on, 2015. IEEE, 2707-2714.

Sohn, T., Varshavsky, A., Lamarca, A., Chen, M. Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W. G. & De Lara, E. Mobility detection using everyday GSM traces. International Conference on Ubiquitous Computing, 2006. Springer, 212-224.

Train, K. E. 2009. *Discrete choice methods with simulation,* Cambridge university press.

Transport Systems Catapult 2015. The Transport Data Revolution - Investigation into the data required to support and drive intelligent mobility. Milton Keynes: Transport Systems Catapult.

UN Global Pulse 2017. Using Big Data Analytics for Improved Public Transport. Jakarta: United Nations Global Pulse.

Wang, H., Calabrese, F., Di Lorenzo, G. & Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on, 2010. IEEE, 318-323.

Watanatada, T., Harral, C., Paterson, W., Dhareshwar, A., Bhandari, A. & Tsunokawa, K. 1987. The Highway Design and Maintenance Standards Model. *Volume 1 Description of the HDM-III Model.* Baltimore and London: The John Hopkins University Press.

Wong, T. C. J. 2015. *Econometric Models in Transportation.* Ph.D. Thesis, University of California, Irvine.

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. 88th Annual Meeting of the Transportation Research Board, Washington, DC, 2009.

Yuan, Y. & Raubal, M. Extracting dynamic urban mobility patterns from mobile phone data. International Conference on Geographic Information Science, 2012. Springer, 354-367.

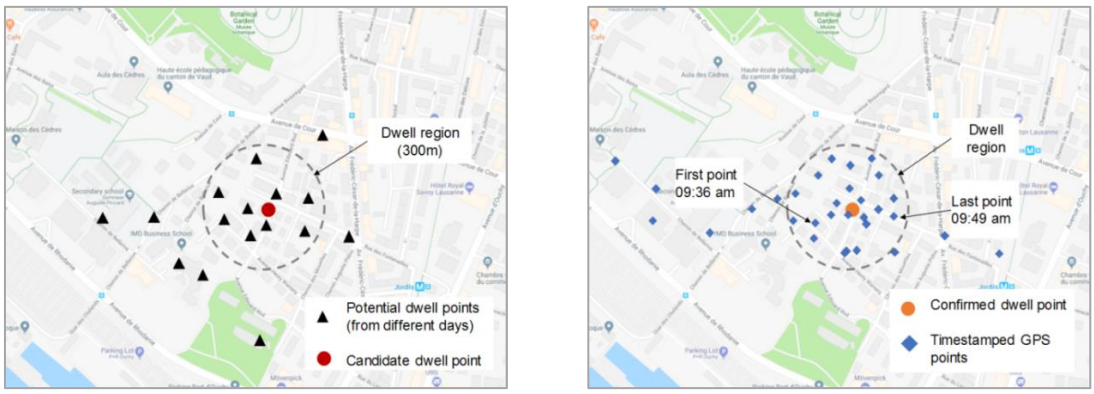
Appendices

Appendix A: Cluster identification for GPS data

As cluster analysis on large datasets is a very challenging task because most spatial clustering algorithms require a full distance matrix, we conducted the clustering in stages. We first split the data of each user according to the date observed. Full distance matrices comprising of all the possible pairs of GPS points observed on a particular day were then generated (Nychka et al., 2018). Thereafter, we conducted complete-linkage hierarchical based on the matrices of each day to identify groups of points that were potentially linked to the same dwell location (Everitt et al., 2011, Murtagh, 1985). Complete-linkage clustering ensures that we constrain the output cluster diameter to a specified size. This is particularly desirable when we know the accuracy range of the records. In this study, we specified a threshold distance of 300 meters as used in previous studies (Çolak et al., 2015, Jiang et al., 2013).

However, it is worth noting that spatial clustering algorithms need at least two data points to identify clusters. As a result, some of the identified clusters might have few points, and would not pass as potential dwell locations. We therefore specified a minimum duration of at least 10 minutes per day calculated using consecutive GPS points. The centroids of the points within the identified clusters were then labelled as potential dwell points. The potential dwell points of each user from different days were combined and complete-linkage clustering conducted again with a threshold distance of 300 meters. This was aimed at clustering potential dwell points from different days in the vicinity of one another, thereby limiting the dwell region size to 300 meters. The centroids of the dwell regions of each user were then computed and labelled as candidate dwell points. At this stage, we did not impose a lower limit on the number of potential dwell points within dwell regions. Therefore, isolated potential dwell points that did not form clusters were simply re-labelled as candidate dwell points.

After establishing the candidate dwell points of each user, we identified all the GPS points within a radius of 150 meters from these locations and ordered the data according to timestamp. This was followed by applying a minimum dwell time constraint of 10 minutes each time a user was continuously observed within the vicinity of a candidate dwell point. Whenever this condition was met, the candidate dwell point was relabelled as a confirmed dwell point. This is illustrated in Figure A1.



(a) (b)

Figure A1 GPS dwell point identification (a) identification of a candidate dwell point from the potential dwell points, (b) application of a dwell time constraint to confirm the candidate dwell point

Appendix B: Cleaning the trip data to identify travel modes

B1. Setting the minimum travel time constraint

To begin with, it is important to note that the observed travel times of the users relate to the inter-boundary components of the O-D links since the trip start and end times are only captured when the users cross the home/work location dwell boundaries. However, these inter-boundary travel times need to be sufficient to enable the observation of reasonable variations in travel time across different time periods. As earlier mentioned, the morning and evening peak travel time increment factors for Lausanne are 1.44 and 1.63 respectively. Since these factors are quite low, for very close O-D pairs, the variations in travel time would not be significant enough to influence changes in departure time choices. In this study, we specify a median travel time of 10 minutes as the lower threshold for direct trips between the users' home-to-work O-D pairs and only consider those meeting this criterion. It may be noted that the exclusion of close O-D pairs also mitigates the observation of potential false trips due to signal jumps that were undetected during the data pre-processing phase (Iqbal et al., 2014).

B2. Identification of trips with unreasonably long travel times

We analyse each user's travel time for a particular trip in relation to the minimum travel time observed for the user along the same trip chain to identify trips with unreasonably long travel times. Travel times generally increase due to traffic congestion, however, when the increase is very big, we suspect other factors such as uncaptured trip segments due to switching off of phones.

To determine the most reasonable upper limits of travel time, we calculate the ratios of the observed travel times versus the minimum travel times for each of the user's trips. These ratios give an indication of the levels of congestion (i.e. the higher the ratio, the higher the level of congestion). We then combine the computed ratios for all the users and estimate the upper limit as follows; $Upper\ limit = Q3 + 1.5 * (Q3 - Q1)$, where $Q1$ and $Q3$ are the first and third quartiles respectively (Tukey, 1977). We use the GSM data for this analysis as it captures most of the trips made.

The estimated upper limits of the ratios were 2.21 and 2.12 for the home-to-work, and the work-to-home commutes respectively. It may be noted that these limits seem reasonable when compared to the congestion factors reported for Lausanne, that is, 1.44 and 1.63 for the morning and the evening peaks respectively (TomTom, 2016). We exclude trips whose travel times exceeded the estimated upper limits.

B3. Identification of potential travel modes

We first apply a minimum distance constraint of 5 kilometres as previous studies have shown that people are less likely to walk or cycle beyond this distance (Hydén et al., 1999). It may be noted that we do not use euclidean distances, rather, we calculate the minimum distances by road for each O-D pair using the Google Distance Matrix API (Google Developers, 2018).

However, another important aspect is the speed of the users. We only consider trip chains where the users' median speeds exceed 15 kilometres per hour, thereby excluding those where the users typically walk or cycle (Bernardi and Rupi, 2015). It may be noted that the calculated speeds are generally over-estimated since we use centre-to-centre O-D distances versus the inter-boundary travel times. Despite this limitation, observing median speeds above 15 kilometres per hour for trip lengths above 5 kilometres is considered a good indicator that the users generally use motorised transport for those trip chains.

References

Bernardi, S. & Rupi, F. 2015. An analysis of bicycle travel speed and disturbances on off-street and on-street facilities. *Transportation Research Procedia,* 5**,** 82-94.

Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiretta, S. R. & González, M. C. Analyzing Cell Phone Location Data for Urban Travel: Current Methods, Limitations and Opportunities. Transportation Research Board 94th Annual Meeting, 2015.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. 2011. Hierarchical clustering. *Cluster Analysis, 5th Edition***,** 71-110.

Google Developers. 2018. *Distance Matrix Service* [Online]. Google. Available: https://developers.google.com/maps/documentation/javascript/distancematrix [Accessed 29 June 2018].

Hydén, C., Nilsson, A. & Risser, R. 1999. How to enhance WALking and CYcliNG instead of shorter car trips and to make these modes safer. Public. Deliverable D6. Walcyng Contract No: UR-96-SC. 099. Department of Traffic Planning and Engineering, University of Lund, Sweden & FACTUM Chaloupka, Praschl & Risser OHG, Vienna, Austria.

Iqbal, M. S., Choudhury, C. F., Wang, P. & González, M. C. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies,* 40**,** 63-74.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira Jr, J., Frazzoli, E. & González, M. C. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013. ACM, 2.

Murtagh, F. 1985. Multidimensional clustering algorithms. *Compstat Lectures, Vienna: Physika Verlag, 1985*.

Nychka, D., Furrer, R., Paige, J. & Sain, S. 2018. *Package 'fields'*, The Comprehensive R Archive Network (CRAN).

TomTom. 2016. *Tomtom Traffic Index - Measuring Congestion Worldwide* [Online]. TomTom International BV. Available: https://www.tomtom.com/en_gb/trafficindex/list?citySize=ALL&continent=ALL&country=CH [Accessed 26 May 2018].

Tukey, J. W. 1977. *Exploratory data analysis*, Addison-Wesley.