Modelling individual accessibility using Bayesian networks: A capabilities approach

Athanasios Bantis
Department of Civil, Environmental and Geomatics Engineering
University College London
A thesis submitted for the degree of
Doctor In Engineering

Declaration

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

Abstract

The ability of an individual to reach and engage with basic services such as healthcare, education and activities such as employment is a fundamental aspect of their wellbeing. Within transport studies, accessibility is considered to be a valuable concept that can be used to generate insights on issues related to social exclusion due to limited access to transport options. Recently, researchers have attempted to link accessibility with popular theories of social justice such as Amartya Sen's Capabilities Approach (CA). Such studies have set the theoretical foundations on the way accessibility can be expressed through the CA, however, attempts to operationalise this approach remain fragmented and predominantly qualitative in nature.

The data landscape however, has changed over the last decade providing an unprecedented quantity of transport related data at an individual level. Mobility data from different sources have the potential to contribute to the understanding of individual accessibility and its relation to phenomena such as social exclusion. At the same time, the unlabelled nature of such data present a considerable challenge, as a non-trivial step of inference is required if one is to deduce the transportation modes used and activities reached.

This thesis develops a novel framework for accessibility modelling using the CA as theoretical foundation. Within the scope of this thesis, this is used to assess the levels of equality experienced by individuals belonging to different population groups and its link to transport related social exclusion.

In the proposed approach, activities reached and transportation modes used are considered manifestations of individual hidden capabilities. A modelling framework using dynamic Bayesian networks is developed to quantify and assess the relationships and dynamics of the different components influencing the capabilities sets. The developed approach can also provide inferential capabilities for activity type and transportation mode detection, making it suitable for use with unlabelled mobility data such as Automatic Fare Collection Systems (AFC), mobile phone and social media. The usefulness of the proposed framework is demonstrated through three case studies.

In the first case study, mobile phone data were used to explore the interaction of individuals with different public transportation modes. It was found that assumptions about individual mobility preferences derived from travel surveys may not always hold, providing evidence for the significance of personal characteristics to the choices of transportation modes. In the second case, the proposed

framework is used for activity type inference, testing the limits of accuracy that can be achieved from unlabelled social media data. A combination of the previous case studies, the third case further defines a generative model which is used to develop the proposed capabilities approach to accessibility model. Using data from London's Automatic Fare Collection Systems (AFC) system, the elements of the capabilities set are explicitly defined and linked with an individual's personal characteristics, external variables and functionings. The results are used to explore the link between social exclusion and transport disadvantage, revealing distinct patterns that can be attributed to different accessibility levels.

Impact statement

The ability of individuals to reach and engage with basic services such as healthcare, education and employment is a fundamental aspect of their well-being. Sustainable development for all implies fair access to transport and to the available destinations and opportunities being offered. Within transport studies, accessibility is considered to be a valuable concept that can be used to generate insights on issues related to fairness and equality due to limited access to transport options. Popular theories of social justice such as Amartya Sen's Capabilities Approach (CA) can contribute towards conceptualising accessibility, allowing thus for a more complete evaluation of issues of transport related social exclusion. In this thesis, implementation of this relationship was done through a coherent model using probabilistic graphical models in general, and dynamic Bayesian networks in particular.

This thesis benefits academic literature in a number of ways. Within transport geography, a novel data driven/graphical model approach to accessibility is introduced using the CA as a theoretical framework. The usefulness of the proposed framework is demonstrated by assessing the equality levels and their link to transport related social exclusion of different population groups in London, using unlabelled, service provider generated mobility data. The proposal and findings have been published in Journal of Transport Geography 84 (2020). In this way, the scope of accessibility appraisals is broadened, retaining at the same time the focus on the individual. Within transportation research literature, the journey to model development made notable contributions in the fields of transportation mode and activity type detection from low resolution mobility data using dynamic Bayesian networks. Two publications have been produced within this field, in the journal Transportation Research Part C: Emerging Technologies 80 (2017) and ISPRS International Journal of Geo-Information 8.12 (2019).

Outside academia, the proposed modelling framework has the potential to provide decision makers with the information needed to assess specific barriers and enablers for each accessibility component at an individual level. Citizens can then benefit from better access to different activity types, improving their quality of life and overall well-being. From a service provider's point of view, the proposed model can provide city and transport planners with mobility and activity patterns of individuals, accounting for characteristics of the environment and stratified by individual sociodemographic characteristics. Such information can be used to support decisions related to the expansion of existing transportation networks or identification of new areas for development. This is an improvement over the

traditional models used within the field of urban planning which tend to focus more on aggregated flows. Within this setting, the methodologies developed in this thesis were applied in the context of transportation mode detection for the Brazilian city of Belo Horizonte. The findings were used by the city's transport service provider (BHTrans) to generate insights related to the quality of reaching and interacting with the transport services from a user's perspective, in a joint project between the Foreign Commonwealth Office, Future Cities Catapult and BHTrans.

Acknowledgements

This work could not have been completed without the help and guidance of my main supervisor Dr. James Haworth, thank you James. I would also like to thank everyone who has been involved in this thesis, particularly Prof. John Twigg and Dr. Catherine Holloway. Further, i would like to thank TfL for the data provision and their enthusiasm about the project.

Finally, the journey towards this thesis would not have been the same without the support of the loves of my life, Annie and Rae.

Contents

Glossary

AFC Automatic Fare Collection Systems.

ANN Artificial Neural Networks.

API Application Programming Interface.

AUROC Area Under Receiver Operating Curve.

BN Bayesian Networks.

CA Capabilities Approach.

CAA Capabilities Approach to Accessibility.

CDR Call Detail Records.

DBN Dynamic Bayesian Networks.

DT Decision Trees.

FCA Floating Catchment Area.

GLM Generalised Linear Models.

GPS Global Positioning System.

GTFS General Transport Feed Specification.

HMM Hidden Markov Models.

IMD Index of Multiple Deprivation.

LDA Latent Dirichlet Allocation.

MNL Multinomial Logit Model.

OD Origin Destination.

PCA Principal Component Analysis.

PGMs Probabilistic Graphical Models.

POI Points of Interest.

PPA Potential Path Area.

PTAL Public Transport Accessibility Levels.

RF Random Forests.

SEM Structural Equation Models.

SIMs Spatial Interaction Models.

STP Space Time Prism.

SVM Support Vector Machines.

UGM Undirected Graphical Model.

Chapter 1

Introduction

1.1 Accessibility in transportation studies

A fundamental aspect of individuals' well-being is the ability to reach the available activities and engage with the opportunities one values. This statement has been used by a range of different disciplines to reflect the notion of accessibility. Within the transportation modelling discipline, accessibility expresses the ability of an individual to make use of the transport system and is closely related to infrastructure characteristics such as proximity to access points, reliability and frequency of the transport system as well as cost (Fransen et al. 2015). Within the discipline of geography, accessibility refers to the extent to which people can access employment opportunities or goods/services given their location and available transportation options (Ettema & Timmermans 2007).

In each case, limited access to both transport options and available opportunities has long been recognised to contribute to the phenomenon of social exclusion (Social Exclusion Unit 2003). Social exclusion as a term has been defined in many ways in the literature but it generally involves the inability to participate in all activities deemed normal in a society as a result of lack of resources, rights, goods and/or services (Levitas et al. 2007). In this context, people who experience limited access to different transport modes are at higher risk of experiencing exclusion from opportunities.

Within the broader definition of accessibility the words "ability", "reaching" and "opportunities" stand out. The abilities of different individuals to use the available transport modes are not the same, and depend on a myriad of factors ranging from the individual's background, socio-demographic characteristics and individual preferences to name a few. Likewise, the acts of reaching the available transport modes and opportunities at a destination are not isolated from the wider physical and socioeconomic environment. Place based characteristics of the

built environment, such as the availability of existing destinations at particular locations, can potentially shape the levels of access to important daily activities such as education and health, employment, leisure and shopping. Constraints such as available time budget and geographical location are also recognised as important factors in the literature, shaping an individual's access to opportunities (Church et al. 2000, Kwan 1998). All these factors tend not to appear in isolation, but to interact with each other in an evolving manner.

The complex and multidimensional nature of the term has made modelling of accessibility extremely challenging, leading to different modelling formulations depending on the research goal, the disciplinary focus and data availability. These range from indicators related to the levels of service for infrastructure assets within an area (e.g. duration of travel and number of opportunities reachable within a specified distance) (Geurs & Ritsema van Eck 2001) to more complicated measures that take into consideration the relative attractiveness of destinations and the benefits provided from a spatial choice set (Miller 2005). The modelling formulations follow from assumptions rooted in different theories, such as utility theory (Ben-Akiva 1979) and time geography (Hägerstraand 1970). These assumptions dictate the level of interactions between the different accessibility components. However, representing the combination of personal abilities and the wider socioeconomic and built environment in a way that reflects the capability of a person to reach potential opportunities remains a challenge (Tyler 2006, Pereira et al. 2017). This is particularly true considering that each component interacts with accessibility in different ways. For example, the socioeconomic background of an individual can have a lasting influence in the ability to use different transport modes and reach different opportunities. On the other hand, place based characteristics, such as the availability of transportation modes, can vary depending on the geography of an area and the time of day. Modelling the relative importance of these interactions requires a flexible specification that can include a wide range of accessibility related assumptions while at the same time accounting for the dynamic nature of the different accessibility components, which is a challenging task given the functional forms of traditional accessibility models. This is increasingly important within the context of studying transport related social exclusion, where an emphasis on the causal structure and interactions of the different accessibility factors is essential (Lucas 2012). For example, even with a fully accessible transportation network, people with disability would experience social exclusion if the destinations lack accessible facilities.

Recently, there has been a growing body of research suggesting that the relation between accessibility and transport related social exclusion can be conceptu-

alised through Amartya Sen's Capabilities Approach (CA) (Preston & Rajé 2007, Hickman et al. 2017, Pereira et al. 2017). At its core, the CA is a theory of justice stating that capabilities are a set of opportunities and freedoms people can choose to act upon (Sen 2008). Viewing accessibility as a basic human capability and using the CA as a theoretical foundation, the components that influence accessibility can be explicitly defined, and the relationships between personal characteristics/wider environment (physical and social) and capabilities (the potential of using transportation modes and the potential of accessing opportunities) can be effectively represented. Extending the modelling context both spatially and temporally to consider location and time (Farrington & Farrington 2005), would enable representation of the evolving nature of an individual's capabilities. As Pereira et al. (2017) puts it, most of the existing accessibility measures are not compatible with this approach and the ones that potentially are, such as person based accessibility measures, fail to draw conclusions beyond the limits imposed by traditional transport surveys.

Survey based studies such as travel surveys and travel diaries (via web, phone, interviews alone or coupled with Global Positioning System (GPS) loggers) have undoubtedly dominated research in both accessibility and transport geography for very good reasons. Such studies reveal both behavioural (e.g. travel purpose) and travel characteristics (e.g. travel frequency, transportation mode etc.) as well as personal characteristics (e.g. income, age etc.) causes, motivations, opinions and situational responses of individual travellers/households (Wermuth et al. 2003), providing in this way, an in-depth view of individual's mobility behaviour. On the other hand, the limitations of survey studies are well documented in the literature. Sample sizes and response rates decrease while participant dropout rates increase. This has a direct impact on population representativeness, particularly for disadvantaged population groups (Stopher & Greaves 2007). Issues of misreporting and omitting important journeys (particularly small journey legs and leisure related trips) or journey attributes are also prominent, affecting the quality of data and modelling outputs (Forrest & Pearson 2005, Wolf et al. 2003). The costly nature of surveys make frequent collections infeasible (Stopher & Greaves 2007), and this is expected to be particularly true for developing countries with limited resources for large scale travel surveys. Finally, capturing the increasingly complex travel behaviour for many individuals is a challenge for traditional travel surveys (Bohte & Maat 2009).

The data landscape however, has changed enormously over the last decade providing an unprecedented quantity of transport related longitudinal data at an individual level. Mobility data from smartphones, automatic fare collection

systems and social networks have the potential to contribute to the understanding of individual accessibility and its relation to phenomena such as social exclusion (Chen et al. 2016, Van Wee 2016). Leveraging such data would not only make possible a recurring, timely and cost effective evaluation of accessibility with a near real-time potential (as opposed to snapshot evaluations dictated by travel surveys), it would also enable the assessment of the relative differences between individuals with different characteristics. This is due to the high spatial and temporal resolution of passively generated mobility data, allowing characteristic patterns of accessibility to emerge.

The potentials of machine generated mobility data however, are coupled with considerable challenges. The quantity of such data does not necessarily compensate for the quality of information contained in traditional travel surveys. The opportunistic nature of data generated by service providers raise issues of population representativeness and bias, limiting the scope of accessibility evaluations in terms of population groups covered, activities reached and transportation modes used (Witlox 2015). Furthermore, modelling attempts raise important methodological challenges (Van Wee 2016). These refer to the unlabelled nature of passively generated mobility data, which often require a non-trivial step of inference if one is to deduce elements such as transportation modes used and activities reached. Moreover, the levels of spatiotemporal resolution of machine generated mobility data can be low, depending on the needs of the service providers. These issues add a computational intelligence aspect to the modelling framework, and introduce an element of uncertainty originating from modelling and data limitations, both of which need to be considered in the accessibility modelling framework.

Clearly, putting the limitations of both data sources aside, a combination between the quantitative breadth of machine generated data and the qualitative depth of travel surveys is the golden medium that could enable facilitation of a more comprehensive approach to accessibility modelling, structured around theories of social justice such as the CA. Not only would such a combination allow linking detailed mobility patterns with the socio demographic profile of individuals, expanding thus the scope of equity evaluations, but it would also inform the modelling of a more detailed individual spatiotemporal behaviour. In this way, assumptions related to the nature of qualitative characteristics of individuals could be readjusted in the face of evidence from machine generated data, and the modelling of individual accessibility behaviour could be informed in the face of individual socio demographic data.

Based on the discussion in the preceding paragraphs, a number of requirements can be defined for an accessibility modelling framework geared towards studying

transport related social exclusion: a) Conceptualising accessibility using the CA as the basis for investigating equity issues in transport requires expressing accessibility components and their interactions in a hierarchical, structured way that enables statistical reasoning; b) Attempting to benefit from the advantages of both machine generated mobility data and information contained in traditional travel surveys requires a modelling framework that enables the combination of diverse sources of data; c) Extracting high level accessibility components from low level mobility data (transportation modes used and activity types reached) requires a computational intelligence capability embedded in the modelling framework; d) Given the different levels of uncertainty related to both data limitations and modelling processes, there is a requirement for a modelling approach that can quantify and make explicit uncertainty related statements. As Martens (2015) puts it, current accessibility indicators fail to capture all these requirements, particularly for applications related to social justice in transport.

1.2 Challenges in measuring accessibility

Literature on measuring accessibility has made considerable advancements in both methodological approaches and on the theoretical link with social issues (e.g (Rasouli & Timmermans 2014, Kwan & Kotsev 2015, Kamruzzaman et al. 2016, Pereira et al. 2017). However, several challenges still remain. In particular:

How should accessibility be defined? What should a model of accessibility include?

Although accessibility as a concept is over three decades old, there is no consensus among researchers, practitioners and policy makers on a definition and how it should be approached in an applied way. The major difficulty lies in the all-encompassing nature of the term which covers diverse topics such as: assessing the influence of different types of barriers to reaching an activity, providing insights into transport disadvantage as experienced by different population groups, as well as the different levels of expressing accessibility in terms of geographies covered, temporal restrictions, modes of egress considered, types of activities as well as research units (eg. service provider, geographies, people) (Geurs & Van Wee 2004). Furthermore, considering the link of accessibility with complex social issues such as freedom of choice (Van Wee 2011), justice (Pereira et al. 2017) and social exclusion (Kamruzzaman et al. 2016), it is nearly impossible to capture all factors that best represent these concepts, both from a theoretical and a computational tractability

point of view. A measurement model using the CA as a theoretical framework offers some advantages in this regard. First, it can be used to identify the most important factors of accessibility as well as the levels of interaction between them. This simplifies modelling as the state space of potential variables and interactions can be reduced to the ones deemed important according to the causal structure imposed by the CA. Second, the latent nature of capabilities in the CA can be used to expand the scope of information that can be included in the modelling process. This is done by using the parameters of the latent capability variables to encode additional information in the model. Finally, expressing accessibility using a theory of social justice provides a framework for investigating transport related social exclusion and equity. This is done by directing the focus of the analysis to the CA components and assessing how they relate to an individual's potential of using different transportation modes and potential of reaching different activity types.

How can the output of an accessibility analysis be interpreted?

Another challenge in measuring accessibility relates to interpreting the outputs of accessibility analyses. On what basis should the outputs of accessibility indicators be used? Policy makers and urban planners are accustomed to using absolute metrics such as transport accessibility levels, or even simpler metrics such as congestion levels or travel time/speed (Papa et al. 2017), which are easy to understand and interpret. However, they are often inadequate in identifying relative differences between people (Church & Marston 2003). Indeed, absolute accessibility measures can obscure people's individual accessibility levels, which can vary greatly depending on their capabilities and choices (Farrington & Farrington 2005). These differences can be crucial to provide evidence about issues such as social exclusion due to lack of transport availability or opportunities at a destination.

Accessibility indicators focused on the individual, on the other hand, are considered to be well suited for providing an in-depth view on the mobility dynamics of individuals (Neutens et al. 2011). However, they are commonly criticised on the grounds of providing a descriptive, as opposed to inferential, view on individual accessibility (Páez et al. 2010). This holds true for both space-time and utility based accessibility measures since both of those modelling approaches focus on the actual travel behaviour of individuals rather than the possible potential behaviours (Martens & Golub 2011). Moreover, current accessibility measures conceal the complex interactions of accessibility factors by condensing the output

to one indicator (Martens 2015, Kamruzzaman et al. 2016). This, although beneficial to decision makers and planners, makes interpretation of the relative influence of accessibility factors difficult. Furthermore, since each measure addresses different accessibility components, the resulting outputs can be considerably different from one another (Neutens, Schwanen, Witlox & De Maeyer 2010). These issues limit their use as relative measures that can be used to quantify differences in individual accessibility levels. Structuring accessibility using the CA as a theoretical basis and elaborating on its basic components (latent capabilities, realised functionings and functioning vectors) through the use of probability distributions over the potential transport modes an individual can use and the potential activity types an individual can reach, could provide the basis for relative accessibility evaluation that can be used regardless of any wider evaluation frameworks (such as cost-benefit analysis) (Van Wee 2016).

What modelling approach should be used?

This challenge relates to the modelling approach upon which accessibility studies are based. There is a long history within accessibility measuring literature that illustrates the evolving nature of the topic. This ranges from using a physical analogy to gravity (Hansen 1959) and mathematically formalising it as a function using the degree of attractiveness and distance (Weibull 1976), to the notion prescribing that activities are both spatially and temporally constrained (Hägerstrand 1973), to conceptualising accessibility based on random utility theory principles (Ben-Akiva 1979). The mathematical implementation of each measure reflects the assumptions used in the modelling process. However, none fully encompases the implementation requirements set out in section 1.1. These are: Representing and quantifying the interactions between different accessibility components in a structured and flexible way; Combining different sources of data, both opportunistic and survey based; Providing computational inferential capabilities for extracting semantic information from unlabelled data; Robustly quantifying the uncertainty in the estimates.

Geometric accessibility constructs such as space-time measures are not particularly suited for quantifying the magnitude and significance of interactions between accessibility factors. On the other hand, utility based accessibility measures assume a linear additive function of accessibility factors, which makes capturing complex interactions challenging. Furthermore, in terms of data requirements, none of the existing accessibility measures are designed to extract high level semantic information from low level unlabelled mobility data (such as transportation

modes used and activity types reached) while retaining the flexibility of combining both machine generated and traditional travel survey data. This is important in the context of developing a model that would enable decision makers to benefit from the longitudinal, pervasive and recurrent nature of service provider data. As a result, a modelling approach that can address the requirements set out in the introduction of this thesis would be beneficial.

To this extent, of particular interest for accessibility modelling is Probabilistic Graphical Models (PGMs). This family of graphical models has been underexplored in accessibility literature, despite their potential for accessibility studies (Kwan et al. 2003). PGMs are capable of expressing complex relationships, including both latent (potential/hidden) and observed (actual/realised) variables, dictated by a graphical structure. Adopting a Bayesian approach, PGMs can facilitate the inclusion of diverse data sets at different levels in the modelling hierarchy to assist inference and expand the information context of the model. In the field of pattern recognition and classification using unlabeled mobility data, PGMs have been found to perform well even with data of low spatiotemporal resolution (Lin & Hsu 2014). The flexibility of PGMs allows straightforward extension to the temporal domain while retaining the overall structure. Finally, the generative modelling framework of PGMs is centered around probability distributions, which enables uncertainty statements about the state of belief of variables through the posterior credible intervals (Koller & Friedman 2009).

1.3 Research aims and objectives

Following the discussion above, the overarching aims of this thesis can be condensed to:

- Developing a novel modelling framework for expressing individual accessibility using the CA as a theoretical foundation, using a combination of travel survey and machine generated data.
- Understanding the differences in accessibility between individuals belonging to different population groups.
- Provide evidence on the levels of equality experienced by individuals, as well as their link to transport related social exclusion.

The specific objectives of this thesis towards the achievement of the above aims are:

- To translate the modelling framework into a quantitative model using PGMs in general, and Dynamic Bayesian Networks (DBN) in particular.
- To perform transportation mode and activity type inference on unlabelled mobility data of different spatial and temporal resolution, quantifying the levels of achievable accuracy.
- To capture the dynamics of an individual's interactions with public transportation modes and available activity types.
- To elaborate on socioeconomic and place based factors that result in the observed levels of transportation mode used and activity types reached.
- To use the inferred capabilities sets for activity types and transportation modes within an entropy based equality assessment framework.

The above objectives will be explored using different machine generated mobility datasets:

- Mobility data gathered by smart-phone devices.
- Social networks mobility data.
- Transportation service provider data (Automatic Fare Collection data).

In particular, the first dataset is used to capture the mobility patterns of individuals experiencing different mobility impairments. In this way, the computational intelligence capabilities of the model in terms of transportation mode detection are demonstrated, introducing at the same time a data fusion framework of travel survey and machine generated mobility data.

Recognising that semantic information on the nature of activities at a destination is an indispensable part of accessibility modelling, the second dataset is used to perform activity type inference while at the same time, benchmarking the limits of accuracy of activity type inference using machine generated mobility data.

Finally, by consolidating the modelling approaches and outputs of the previous models, a novel framework of expressing accessibility at the level of an individual based on the basic elements of the CA is introduced. The usefulness of the proposed framework is demonstrated using the third dataset, where the link between equality levels and transport related social exclusion of different population groups is assessed.

1.4 Scope of the thesis

Although rural accessibility is a very important topic of research in its own right, this thesis will concentrate only on urban environments, with its geographical scope being the Greater London area. The model formulation is always built from the ground up, focusing on individual accessibility. Contrary to the majority of accessibility studies or studies using the CA in transport, this thesis uses a probabilistic approach to modelling mobility and activity patterns from unlabelled mobility data. In this way it differs from studies using travel diaries and studies using qualitative data from interviews and focus groups, where the activities performed and transportation modes used are explicitly stated by the participants.

Finally, any accessibility related conclusions, particularly in the later chapters of the thesis (chapter 7) are only relevant to the extent the mobility data used allow. Highlighting this is of particular importance considering the fact that semantic information on people's mobility and accessibility patterns is missing and has to be imputed from secondary data. Nevertheless, as demonstrated in the case studies of this thesis, the developed modelling framework is flexible enough to account for data of different spatial and temporal resolution, and as such, confidence in the results is expected to improve with data of greater fidelity.

1.5 Structure of the thesis

This thesis is organised in 8 chapters. Chapters 1- 4 set the scene by reviewing the relevant literature on accessibility, CA and the relationship between them, as well as providing some background on graphical models. In particular, chapter 2 provides a literature review on the different ways accessibility has been approached, along with a review of the different methodological frameworks that have been used.

Chapter 3 explores the link between accessibility and CA as it has been explored in the transport context. The literature reveals the usefulness of the CA in three regards: 1) as a policy evaluation tool for transport interventions; 2) as a way to frame and extent the notion of accessibility and; 3) as a way to examine the relative transport disadvantage and social exclusion experienced by different population groups. The chapter then examines the potential of using existing accessibility measures with a CA framework before proposing a formulation using graphical models.

Chapter 4 reviews different types of graphical models and assesses their applicability within accessibility and CA studies. DBNs are identified as the most

promising graphical model in terms of meeting the objectives of this thesis.

Chapter 5 introduces the data and chapter 6 describes the methodology used in this thesis. This is defined by three interrelated DBNs, each performing a specific task: transportation mode detection, activity type inference and accessibility modelling structured around the CA (Capabilities Approach to Accessibility (CAA) model).

Chapter 7 presents the modelling results and proceeds to examine the levels of social exclusion and transport disadvantage experienced by some individuals using a popular entropy driven equality index (Theil index) on the posterior quantities of the capabilities sets.

Finally, chapter 8 provides the conclusion of the thesis. The main outcomes are stated and future directions are discussed.

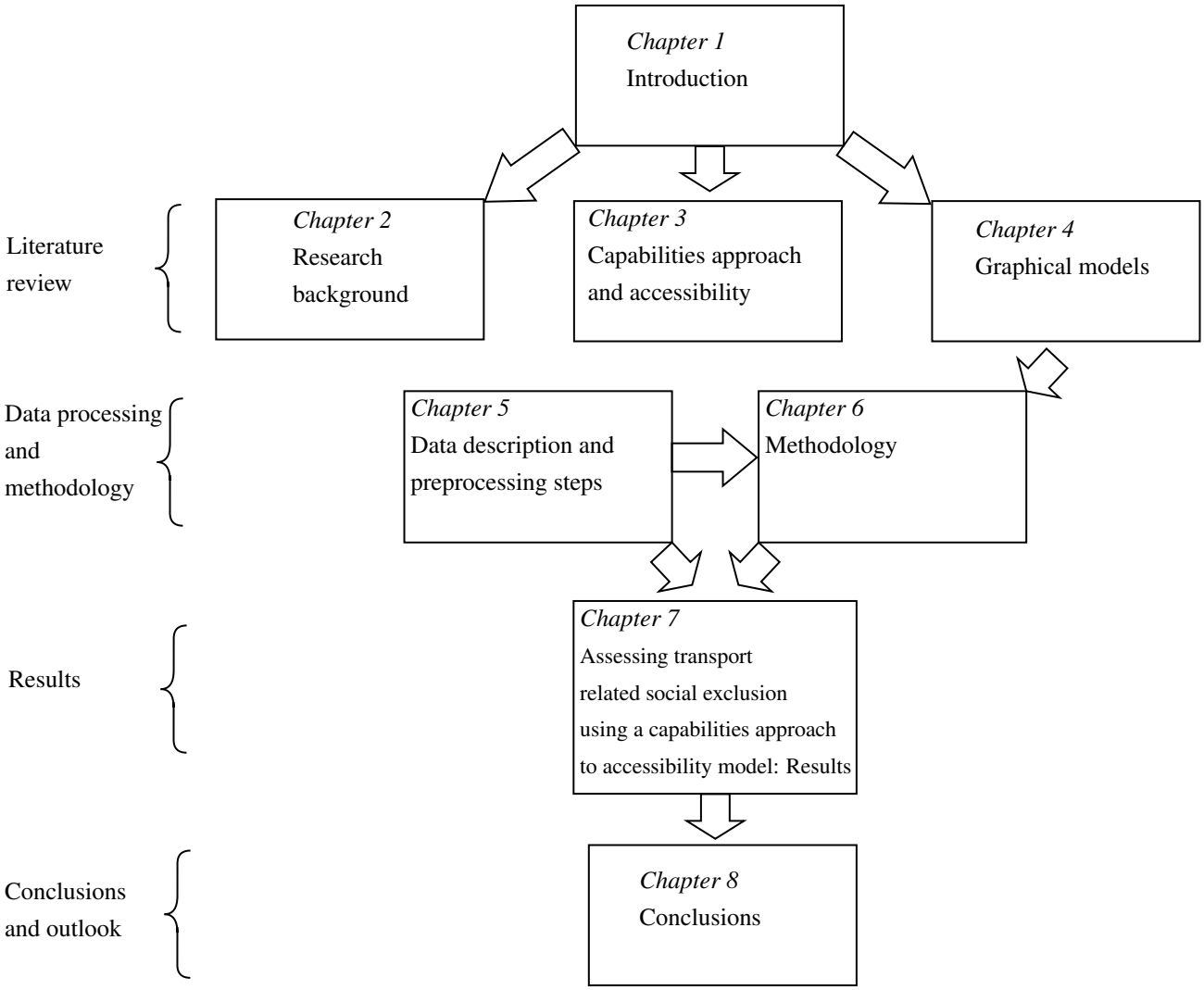


Figure 1.1: Thesis roadmap.

Chapter 2

Research background

2.1 Chapter overview

This chapter provides the research background for this thesis. The roadmap that guides the structure of the literature review is illustrated in the following figure:

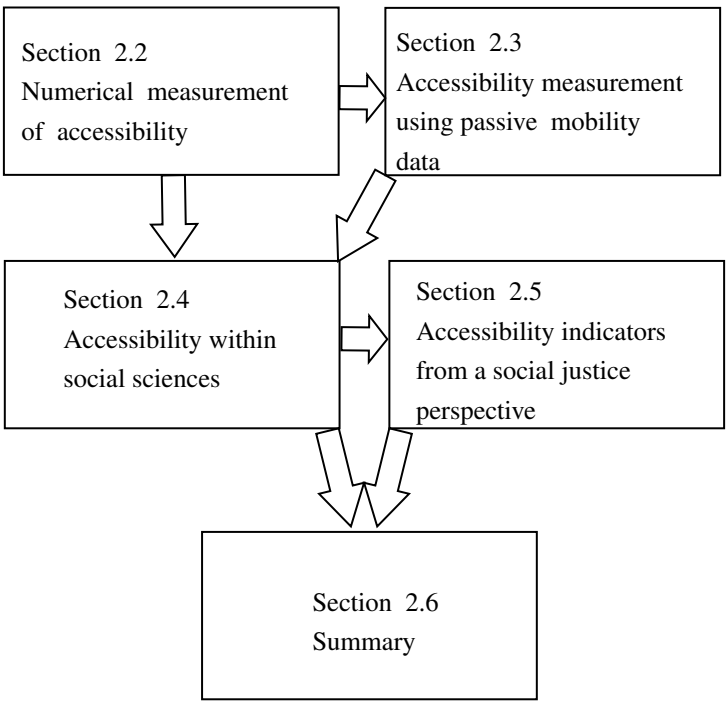


Figure 2.1: Research background roadmap.

Section 2.2 provides an overview of the most widely used accessibility indicators along with their strengths and weaknesses. The focus here is to examine how the different accessibility components are included in the functional forms of the indicators and the implications this has for evaluating equity issues in transport.

Section 2.3 provides a brief overview of the ways passive mobility data have

been used within the context of those indicators, along with the challenges introduced by the unlabelled nature of such data. Relevant literature related to the definition of accessibility and the link between accessibility indicators, equity and transport related social exclusion is presented in section 2.4.

Section 2.5 provides an overview of how different theories of social justice have approached accessibility and the role of the different accessibility indicators for the purposes of equity evaluations. The premise here is that approaching accessibility modelling from a justice theoretic lens is beneficial in that it provides a framework for identifying and structuring the most important components for investigating equity issues in transport.

Consolidating the findings, section 2.6 provides a summary of the most important learnings for the different accessibility families of indicators with respect to theoretical basis, technical, practical and equity related considerations for this thesis, as well as indicative applications.

2.2 Numerical measurement of accessibility

The concept of accessibility has been the focus of different disciplines such as geography, urban planning and transport planning for some time. The wide adoption of the term has resulted in different definitions commonly encountered in the literature: A very early definition originates from Hansen (1959) who defined accessibility as the potential of interaction between destinations. Within transport economics Ben-Akiva (1979) based the definition of accessibility on the benefits provided by the interaction between transport and land use. In transport geography, Geurs & Van Wee (2004) defined accessibility as the extent to which transport and land-use systems enable individuals or groups of individuals to reach activities or destinations by means of transport. These definitions are characterised by different mathematical models and data specifications, applied within different contexts and different measurement levels. This section provides a description of the different accessibility indicators, focusing on their strengths and weaknesses.

2.2.1 Accessibility models

2.2.1.1 Gravity-based indicators

This family of models perceives accessibility as the potential of interaction between different spatial entities (Hansen 1959). In general, these spatial entities are geographical areal units within which a measurable quantity is observed. Examples include measures of populations, number of commuters, number of jobs etc. The

model considers the potential of interaction between an area A and another area B to be proportional to the magnitude of activity potential of area B, and inversely proportional to some impedance function between areas A and B. This definition is closely related to gravity models, a mathematical formulation that is analogous to the Newtonian gravitational law applied to human behaviour:

$$M_{ij} = g(A_i, B_j)/f(c_{ij}) \tag{2.1}$$

where M_{ij} is the potential of interaction between areas ij, B_j is the measure representing the activities/opportunities of area/zone j in relation to the area of zone A_i and cij is the generalised cost between these areas. This cost is usually mapped to a function $f(c_{ij})$. This function commonly appears as an exponential decay function with a scalar decay constant determining the strength of the influence between the pairwise quantities or as a simple power function with the exponent determining the degree of interaction. The function $g(A_i, B_j)$ ensures that the product of A_iB_j is in accordance with the observed quantities, usually through a multiplicative factor.

This model has been extensively used in the field of social sciences for decades, despite its loose theoretical foundation; unlike the accuracy in predictions that Newtonian equations of gravity provide in the physical realm, such results were never empirically established in the field of social studies (Sen & Smith 2012). Nevertheless, the model is popular even today due to its simplicity and intuitiveness.

A very commonly used gravity type model is the potential accessibility measure (Hansen 1959):

$$A_i = \sum_j D_j e^{-\beta c_{ij}} \tag{2.2}$$

where A_i is the potential accessibility of the ith area, D_j is a proxy for the opportunities of area j.

The exponential function implies that nearer destinations (for example, in terms of distance or travel time) to the origin are more accessible than distant ones. The parameter β can have a significant effect on the results and is usually estimated using empirical data of the spatial behaviour of people in each area. Other commonly used decay functions in the literature are Gaussian and logistic functions, however the negative exponential is the most used one as it is more closely related to travel behaviour (Handy & Niemeier 1997).

Applications of this type of measure in real world settings are many and range from assessing the accessibility to jobs, retail, health and education services (Geurs & Ritsema van Eck 2001). In many of the studies, equation 2.2 has been adapted to meet the specific needs of the analysis. For example, Van Wee (2016) and Wegener et al. (2000) have used the logsum impedance function as a generalised cost function for assessing multimodal accessibility. This is defined as:

$$c_{ij} = -1/\beta ln \sum_{m} e^{-\beta_i c_{ijm}} \tag{2.3}$$

where c_{ijm} is the generalised cost between ij for mode m and β is the sensitivity parameter.

Apart from the potential accessibility measure, gravity models formed the basis for developing the concept of potential interaction between spatial entities. This line of research resulted in a family of Spatial Interaction Models (SIMs). The theoretical underpinning of those models is the idea of distance as a friction, which is in accordance with the spatial dimension of accessibility. Besides distance as friction, spatial interaction models assume complementarity and the potential of changing the expected interactions by using some other explanatory variables (Fotheringham & O'Kelly 1989). Complementarity refers to the idea of supply and demand between two areas. For example, an area that has a surplus of jobs is complementary to an area that has high employment demand. Spatial interaction models have extended the simpler gravity models by introducing sensitivity parameters and constraints for both the supply, demand and friction components. These parameters can be formed within both a deterministic and a probabilistic framework (Wilson 1971).

Strengths and weaknesses

Important advantages of the potential accessibility measures are the straightforward communication of the results (this is relevant only for the less complicated gravity type models) and the modest data requirements, as only land use and transport related data are needed (Geurs & Ritsema van Eck 2001, Koenig 1980). Furthermore, such measures are thought to be appropriate as social indicators, as they can be used to analyse the levels of access to different social, healthcare and employment opportunities, particularly the implementations that allow disaggregation based on different population groups (Geurs & Van Wee 2004, Neutens 2015). This is done by assessing the trade-off between the size and quality of the available opportunities (maximising attractiveness) with some form of travel impedance (minimising costs). Given more detailed data such as levels of available

income, it is possible to extend the model to include variations in this trade-off (typically through the sensitivity parameter of 2.2).

It is important to notice the relationship of the functional form for the models belonging in this family with Generalised Linear Models (GLM) such as Poisson or Negative-Binomial regression. This positions the model firmly within statistical estimation literature and makes calibration of the free parameters and balancing factors easier, allowing thus more complicated specifications.

The inverse of balancing factors could be interpreted as an indicator of economic benefit, which allows relating some models of this family of accessibility indicators with aspects of economic theory (Neuburger 1971). At this stage, it is worth mentioning that all gravity type accessibility indicators adhere to a common generalised form, which was formalised as an accessibility measure by (Weibull 1976). In this formulation, accessibility is considered as a property of the spatial configuration of opportunities available for spatial interaction. In this context, a spatial configuration is simply a configuration of opportunity elements $\omega_{1...n}$ belonging to a finite set Ω. Accessibility is then a property of this configuration, materialised as a mapping function that expresses the magnitude of the property. In this axiomatic formalisation, the accessibility indicators preserve some desired properties such as monotonicity, additivity and associativity.

Disadvantages often mentioned in the literature are the propensity for overestimating internal accessibility and the inability to account for competition effects between the supply and demand of each origin/destination pair (Karst & van Eck 2003). The former relates to the self-potential of an origin with considerable mass, that leads to heavy weighting for internal accessibility (Geurs & Ritsema van Eck 2001). The latter implies that the distribution of demand does not affect the accessibility levels of the destinations. Such an assertion is hard to defend since competition often shapes the spatial distribution of activities and can lead to inaccurate and misleading results (Shen 1998). More elaborate models, such as the doubly constrained model, were designed to address these limitations through the inclusion of balancing factors (Geurs & Van Wee 2004, Sen & Smith 2012).

Moreover, although in theory SIMs can be disaggregated to individual population groups, model calibration can be difficult, as different balancing factors have to be computed for all population groups considered. Perhaps the most important disadvantage for the purposes of transport related social exclusion and equity evaluations, is the level of aggregation for these types of models, which is typically done at areal unit level (census tracts, land-use polygons, transportation analysis zones etc.). This is sufficient for applications focused on studying issues of equity of access to different service providers (e.g. primary care Guagliardo (2004),

employment Karner (2018)). However, applications focusing on equity considerations at the level of individual cannot be addressed. This is because individual variations in accessibility cannot be captured at aggregated units of measurement (Dong et al. 2006).

Finally, the non-linear nature of the models (particularly for the most advanced SIMs) and the lack of causal structure between the variables can become a challenge when interpreting the output.

2.2.1.2 Cumulative-based indicators

This measure is also referred to as relative accessibility. It was first proposed by Ingram (1971) and describes accessibility in terms of some form of separation between two points. The separation function can be the physical distance, time or cost. The simplest form of this measure can be considered a straight line connecting two points. Due to its simplicity, this measure has been used as a standard in land-use planning for the distance between a point or an area and the transportation infrastructure (Geurs & Van Wee 2004). The measure can be extended to include multiple destination locations that can be accessed within a given time, distance or cost. The latter is referred to as the isochronic, contour or cumulative approach:

$$M_i = \sum_j d_{ij} W_j \tag{2.4}$$

where d_{ij} is the separation measure and W_j are the destination points.

Figure 2.2 below shows the contours for an origin point in central London using travel time as a separation measure, assuming a mean travel speed of 4.5km/h.

Researchers have proposed improvements over this basic formulation. To address the inability to account for competition effects in the indicator, researchers have proposed different strategies. These range from introducing a distance decay function in the specification (Cheng & Bertolini 2013), to following a probabilistic approach to estimate the share of relative opportunities between the zones defined by the indicator (Kelobonye et al. 2020) and using fuzzy logic to delineate minimum accessibility distance thresholds (Lotfi & Koohsari 2009). Floating Catchment Area (FCA) methods broadly belong in this family of indicators as well. These allow the contours to vary from one location to another using a proxy for the catchment area (e.g. opportunities to population ratio), incorporating assumptions regarding service availability within the contour area (Wang 2000). An extension of the FCA method is the two step FCA (2SFCA) where both the catch-

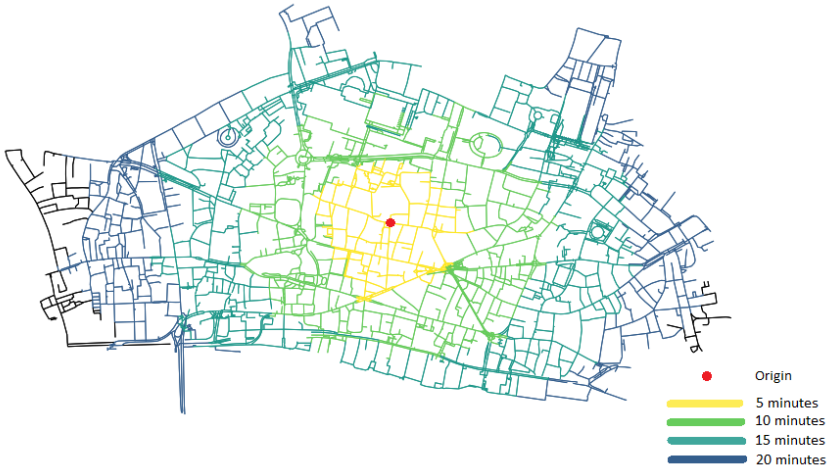


Figure 2.2: Isochrone contours of 5, 10, 15 and 20 minutes walking time from an origin point in the City of London.

ment area for facilities and population is allowed to float based on the location (Delamater 2013).

Strengths and weaknesses

A big advantage of cumulative accessibility measures is their simplicity and intuitive interpretation. This has allowed researchers to easily combine this indicator with other accessibility measures that take into consideration temporal and individual constraints, such as space time accessibility, using map algebra techniques (Fransen & Farber 2019). It is interesting to observe that this indicator has a very similar functional form to gravity type indicators, which makes integration of the two measures straightforward (Páez et al. 2012). Due to their counting property, which makes use of absolute units, and the modest data requirements for indicators of this family, accessibility comparisons between cities is relatively straightforward (Merlin & Hu 2017). This is in contrast to gravity type indicators, where interpretation of the output can only be made in relative terms, after normalisation (Batty 2009). These advantages have made cumulative accessibility popular within the transport equity literature. El-Geneidy, Levinson, Diab, Boisjoly, Verbich & Loong (2016) used the indicator to study the effects of both travel time and transit fares for accessing employment opportunities, focusing on disadvantaged population groups. Pereira (2019) used a travel time based cumulative accessibility measure to study the impact of future transport infrastructure investment scenarios to the accessibility of low income groups. Farber et al. (2014) used a cumulative accessibility indicator along with the General Transport Feed Specification (GTFS) files to investigate accessibility to food options throughout

the day. Fransen et al. (2015) used a similar indicator and data specification in the context of identifying transport gaps and their role in social exclusion. Furthermore, it should be noted that this indicator has been extensively used to study equity issues around access to healthcare (Neutens 2015). This is due to the sensitive nature of patient data where they are seldomly shared among researchers outside clinical research.

Disadvantages of the indicator include the inability to account for attractor variables (at least in its native form), which suggests that all opportunities are equally desirable. Another disadvantage is the arbitrary threshold of isodistance and the lack of sensitivity to the opportunities that are contained within the isochrone (Ben-Akiva 1979). Because of this, Geurs & Van Wee (2004) mention that this measure can lead to misleading decisions during land-use/transport infrastructure projects, as interventions that aim to improve travel time may not lead to an improvement of accessibility. For example, an infrastructure project that reduces the travel time between two points from 50 to 15 minutes, doesn't necessarily improve the accessibility of the destinations within the boundaries of the isochrone polygons. Similar to gravity type indicators, cumulative accessibility measures are aggregate measures, which render them inapplicable for equity evaluations at the level of an individual. However, the approach is still of interest to this research as the basic concepts can be easily implemented within more complex models of accessibility, for example to define the activity space of an individual.

2.2.1.3 Space-time accessibility indicators

Space-time accessibility (also referred to as the constraints oriented approach) traces its roots to the work of Hägerstraand (1970). In this approach, accessibility is evaluated relative to an individual's ability to reach activity locations given constraints such as the person's daily activity schedule and other spatio-temporal constraints (Kwan 1998). In this measure, both the temporal and spatial component are equally important.

The Space Time Prism (STP) is one of most used methods of space-time accessibility measurement. Geometrically, the STP is a construct that defines (Miller 1991):

- the space that can be reached in a given time interval, bounded by the locations of an individual at the begin and end of the journey,
- the time required for participation in activities during that interval and
- the velocities at which an individual can travel.

An illustration of the space-time prism is shown in figure 2.3. In this figure, the individual has to be at a certain location (e.g.. workplace) until time t_1 and then return at that location for time greater than t_2. This leaves time $t_1 < T < t_2$ time units for the person to reach all the available destinations. The level of accessibility using this technique can then be defined by using the size and volume of the space-time prism (potential path space) or its projection on the geographical space (Potential Path Area (PPA)).

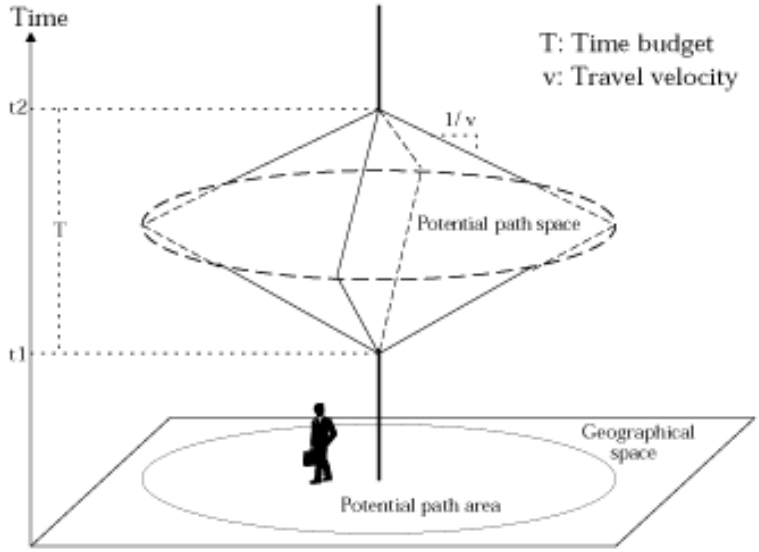


Figure 2.3: Space-time prism. Image source: http://www.rita.dot.gov

Improvements over the basic structure of the STP have also been proposed in the literature. Recognising that there are inherent infrastructure related restrictions of the extent of PPA, Miller (1991) used the road and public transport networks as a means to constrain the area that can be reached (this approach is referred to as network time prism). Wu & Miller (2001) used a dynamic network approach in the calculation of STP that can account for varying travel times due to congestion. Acknowledging the uncertain nature of travel times as well as the start and end time of an individual's time window both in space and time, many of the core components of STP have been extended to include a stochastic element. For example, Kuijpers et al. (2010) extended an individual's STP anchor points (start and end points) to vary according to a set of possible outcomes, each of which has different degrees of uncertainty. Within this line of thought, Chen et al. (2013) proposed an analytical framework to delineate STP under travel uncertainty, ac-

cording to a predetermined arrival time probability. Investigating the day-to-day variations in potential travel and activity behaviour, Neutens, Delafontaine, Scott & De Maeyer (2012) used a dynamic evolution approach to calculating STPs for different individuals. Combining individual based and facility based STPs, Wang et al. (2018) developed a methodology for including restrictions related to service areas and opening times of facilities.

Strengths and weaknesses

In contrast to relative and potential accessibility, space-time accessibility is focused on the household/individual/person level and in this way, equity evaluations at the level of individuals can be performed (Fransen & Farber 2019, Neutens, Schwanen, Witlox & De Maeyer 2010, Kwan 1999). The focus of the indicator on what is reachable given a time budget allows researchers to approach computation of STPs using methods from analytical geometry such as Fourier shape analysis (Lee & Miller 2019). The individual components that describe the STP (eg. PPA) are intuitive and relate directly to an individual's behaviour allowing for a richer representation of a person's capabilities and constraints when reaching an activity (Miller 2016). Moreover, by studying individuals' anchor points (origin and destination points) and durations within the SPT, multipurpose/multimodal assessments of accessibility can be performed (Ettema & Timmermans 2007). Finally, advances in analytical time geography have enabled the construction of SPTs and the study of their theoretical properties using approximation techniques such as Monte Carlo simulations, Random Walks and Kernel Density Estimation techniques (Miller 2016, Kobayashi et al. 2011, Liao et al. 2014)

On the other hand, some authors argue that space-time accessibility measures do not account for competition effects and the capacity constraints of destinations (e.g. available jobs) (Geurs & Van Wee 2004) and as such, they are not suitable for analysing accessibility where competition occurs (e.g. employment). Other authors mentioned that, since space-time accessibility measures focus on short term behavioural patterns, they are not well suited to study the long term effects of land use and transport changes on daily activity patterns (Sclar et al. 2014). Moreover, such measures are difficult to operationalise since they rely on a computationally intensive framework and require extensive and detailed datasets at an individual level (e.g. travel diaries and time use studies). Finally, it has been argued that accessibility evaluation using the space-time approach is framed in terms of the researcher's expectations of an individual's behaviour which does not necessarily reflect the actual observed behaviour. As such, it has been used to indicate individual travel possibilities rather than explicitly predicting or explaining individual

behaviour (Páez et al. 2012, Neutens, Versichele & Schwanen 2010, Pred 1977). In the context of this thesis, reasoning behind the factors that potential explain an individual's accessibility is particularly important, since it enables equity related evaluations between individuals.

2.2.1.4 Utility-based indicators

Utility-based accessibility measures trace their origin in the field of microeconomics and consumer choice theory (Ben-Akiva 1979). In this specification, an individual makes the decision to choose an activity over a discrete choice set, all of which could potentially satisfy an individual's needs (Geurs & Ritsema van Eck 2001).

The utility or value that each individual assigns to a particular choice set is not known to the analyst, so it is common that is treated as a random variable. Traditionally, the utility function of an individual n located at i attaches at a destination j is given by (Geurs & Ritsema van Eck 2001):

$$U_{ij} = V_{ij} + \beta c_{ij} + \epsilon_{ij} \tag{2.5}$$

where U_{ij} is the utility that an individual assigns to the choice set, V_{ij} is the value one gains from a trip, c_{ij} is a general cost function and ϵ is the random error term.

Assuming that an individual will opt to maximise equation 2.5 given a set of choices, then accessibility can be viewed as the configuration of probability choices that maximise this utility function. This corresponds to the denominator of a Multinomial Logit Model (MNL):

$$ln\frac{P(U=\kappa-1)}{P(U=K)} = e^{V_{ij}+\beta c_{ij}+\epsilon_{ij}} \tag{2.6}$$

The observed utility part of the total utility of choice κ is referred to as the log-sum model in accessibility literature (Dong et al. 2006, Geurs et al. 2012, de Jong et al. 2005). Assuming the choice set is $K \in \{1...m\}$:

$$A_i = ln\Big(\sum_{\kappa=1}^{m} e^{V_\kappa}\Big) \tag{2.7}$$

Equation 2.7 can be rewritten relative to the potential accessibility of section 2.2.1.1 by applying the exponential decay function (Geurs & Ritsema van Eck 2001):

$$A_i = \frac{1}{\beta} ln \sum_j D_j e^{-\beta c_{ijm}} \tag{2.8}$$

The second framework of utility-based accessibility measurement is analogous to the doubly constraint version of the spatial interaction model of potential accessibility (Martínez & Araya 2000). In this specification, the utility function of an individual is constrained to comply with the observed origin destination journeys:

$$A_i = \frac{1}{\beta} ln(\alpha_i) \tag{2.9}$$
$$A_j = \frac{1}{\beta} ln(b_j)$$
$$A_{ij} = \frac{1}{\beta} ln(\alpha_i b_j)$$

where α_i are the expected benefits from the trips generated from origin A_i, b_i are the expected benefits from the trips attracted to destination A_j and $\alpha_i b_j$ are the benefits derived from the trip between ij.

Strengths and weaknesses

One important advantage of this family of accessibility indicators is their link with random utility theory as described in microeconomics. This fact, explicitly relates the outputs of the indicator with consumer surplus theory (by dividing equation 2.7 with the cost function). Furthermore, it enables expression of accessibility in monetary terms (Van Wee 2016), opening up the possibility for more complicated model specifications such as discrete choice, nested choice and extreme value models (Ben-Akiva & Lerman 1985). These models can account for some of the shortcomings of the simpler specifications, such as the principle of independence of irrelevant alternatives[1] (IIA) that occur within multinomial logit models (Ben-Akiva & Lerman 1985). This is done by modelling the correlations between the choice sets through partitioning them into more generalised clusters. Examples of such specifications include the nested logit (NL) (Yun et al. 2000), cross-nested logit (CNL) (Ben-Akiva & Bierlaire 1999) and mixed multinomial logit models (MMNL) (De Jong et al. 2003). Versions of such models that use latent variables to capture population heterogeneity have also been proposed, either through a mixture of MNL probabilities or through the use of latent class NL specification (Wen et al. 2012). This is done by dividing the population into a discrete set of classes that can be used to capture the unobserved preferences of individuals using variables such as socioeconomic characteristics (Xiong et al. 2014). At this

[1] A common example to demonstrate the practical effect of IIA is the following: Consider that there is a mode choice of travelling by car or taking a red bus, each having a probability of 0.5. If we introduce a blue bus, IIA tells us that now the probability of taking a bus (blue or red) would be 0.667 which is counterintuitive considering that both buses provide an identical service.

point, it should be mentioned that applications of discrete choice models within transportation literature have a long history, which in the context of parameter estimation, is manifested through reliable and efficient algorithms.

Similarly to space-time accessibility, utility-based accessibility indicators are often classified as individual accessibility indicators and, from this perspective, are well suited to study individual behaviour. Another advantage is that utility-based accessibility scales well, as aggregation of individual measures produces realistic results (Geurs & Ritsema van Eck 2001). All the advantages mentioned above have made this family of indicators attractive for investigating issues of social equity in transport (Neutens, Delafontaine, Schwanen & Van de Weghe 2012, Van Wee & Geurs 2011, Neutens, Schwanen, Witlox & De Maeyer 2010).

On the other hand, a limitation of utility based models is the linear functional form of the value term in the utility function which limits their applicability when modelling more complex individual behaviours (Yamamoto et al. 2002, Xie & Waller 2010, Zhu et al. 2018). This is due to the assumption that all included variables (or predictors) in the model influence the utility of an individual in an additive way, regardless of their place in the modelling hierarchy (as in case of nested or latent variable models). For applications where the process that gave rise to the phenomenon in question is not known or only partially understood (such as the relationship between accessibility and issues around equity), such an assumption can be restrictive. This relates to the issue of how to model an individual's choice process, which in the case of utility based models is deterministic in the predictor variable space. This can be problematic as it is thought that most people reason under "if-then" scenarios rather than using utility maximisation terms (Janssens et al. 2006).

Another issue relates to the correlation structure of multidimensional choice problems. This arises from heterogeneity in preferences which manifests into varying variance across the different choice occasions of the included covariate terms (Keane 1997). This is clearly a problem when the model assumes homogeneity (as in simpler utility based models). Heterogeneity in preferences can also be an issue in more elaborate specifications, as it requires explicit modelling of the specific structure of correlation among the different modelling dimensions, which leads to more complicated model specifications with further assumptions on the structure of the error variance-covariance matrix (Bhat & Guo 2004).

Furthermore, when it comes to behavioural dynamics and interactions between the elements of the choice sets, discrete choice models have almost universally been applied in a static context (Ben-Akiva et al. 2002). Research on dynamic discrete choice models is limited and mainly focuses on simple cases where the

errors between transitions are assumed to be constant and the base model is a binary logit (de Palma & Kilani 2005, De Palma & Kilani 2011, Arcidiacono & Miller 2011). This makes utility based models unsuitable for applications where the evolution of state spaces is important.

Another challenge of utility-based indicators for addressing issues of social equity is the emphasis on an individual's realised behaviour, ignoring the potential or possible behaviours that the person could have chosen if his/her situation was different (Chorus & De Jong 2011, Martens & Golub 2012). Furthermore, utility-based measures are thought to be susceptible to arguments such as the "expensive tastes argument"[2] and the "offensive tastes argument"[3] (Wolff 2007). For example, the utility that is experienced from a buggy user that occupies a wheelchair user spot on a bus should not be accounted as contributing to his/her welfare.

Finally, the data requirements for estimating the choice probabilities for this family of models can be an issue, with the overwhelming majority of studies using survey data where an individual's choice is observed. Apart from making utility-based accessibility evaluations costly and difficult to repeat at frequent intervals, the requirement of observed choice renders these methods inapplicable for unlabeled mobility data where an individual's choice needs to be inferred from contextual information.

2.2.1.5 Other methods

Besides the above described models, researchers have explored other accessibility formulations using different computational constructs.

A popular measure that does not fall under the above described categories is the concept of activity spaces. An activity space is defined by the geographical extent an individual travels for reaching their daily activities. Generally, there are three different approaches for the computation of activity spaces found in the literature (Patterson & Farber 2015). The first is related to the construction of geometric objects centered around a suitably defined point (eg. a public transport access point). These objects can take the form of ellipses, circles or more elaborate constructs such as convex hull polygons. The second approach is related to constructing buffers around the shortest path networks that connect the points visited by the individual. Finally, the third approach is related to the construc-

[2]The "expensive tastes argument" describes a thought experiment stating that, between individuals that have otherwise the same ability to convert resources into welfare, if an individual happens to develop 'expensive tastes' then a distribution of resources from one person to the other is required in order to equalise their resources

[3]The "offensive tastes argument" refers to being denied admission to a good or service on grounds of justice

tion of activity space surfaces by non-parametric methods such as kernel density estimates.

Geometrically defined activity spaces using ellipses and circles have been found to correlate with more simple measures of travel behaviour such as travel distance (Schönfelder 2001). However, since the surface of these constructs is determined only by the spatial distribution of visited locations, they do not consider elements of the environment that could have contributed to the observed mobility behaviour. Kernel density approaches provide more flexible activity space constructs but suffer from the same issues as ellipses and circles, with the additional disadvantage of tending to over-fit to the observed locations. On the other hand, using shortest path network approaches, tends to focus on the connectivity between different visited locations, which may overestimate an individual's activity space for services such as rail.

Other methods include the use of Geographical Information Systems (GIS) in conjunction with the methods described in previous sections to enhance the results and incorporate more assumptions in accessibility computations. For example, Wang & Chen (2015) used a relative job accessibility measure accounting for spatial autocorrelation using a simultaneous autoregressive model (SAR). Outside academia, UK's Department for Transport (Department for Transport 2014) has developed a set of core accessibility indicators to assist local government bodies when undertaking local transport planning. Examples of the indicators include: journey time from origin (O) to nearest destination (X) using mode (Y); Frequency of public transport from O to X; Number of X accessible by X from O in t time; Population of O within t time of X by Y.

2.3 Accessibility measurement using passive mobility data

Traditionally in transportation research, numerical accessibility indicators have been applied using travel surveys and questionnaire surveys (Schönfelder & Axhausen 2003). While the benefits of such surveys are well understood (e.g. including travel purpose, travel frequency, transportation mode etc.) the costly nature, limited sample size, low update rate and low temporal and spatial resolution make them unsuitable for recurrent evaluations of accessibility. Recently, however, there is an increasing number of studies using passive, machine generated mobility data either for computing, or providing data input for accessibility indicators (Martens et al. 2019). As already mentioned in section 1.1 such data have the potential to contribute to accessibility literature due to their low cost, high update rate and

detailed nature. This enables the study of travel behaviour of individuals at a daily trajectory level. In the context of this thesis, passive mobility data refer to data that can be considered opportunistic, usually gathered for purposes other than those of accessibility analysis.

2.3.1 Passive mobility data for calculating accessibility

Commonly used passive mobility datasets used in studies measuring accessibility include Call Detail Records (CDR) data, Automatic Fare Collection Systems (AFC) data, GPS as well as social media data. Some studies have also used data generated using the General Transport Feed Specification (GTFS), however, such data relate to transportation service provider availability, rather than individual mobility behaviour.

In the context of deriving individuals daily activity spaces, Xu et al. (2015) used a mobile phone location dataset to uncover key locations that serve as anchor points, around which everyday activities take place. Using Call Detail Records (CDR) for over 1 million individuals for the city of Shenzhen, China the authors used measures such as radius of gyration and standard distance to describe the spread of activity spaces. The authors found that it is possible to correlate the activity spaces with different economic and transportation characteristics of the study area. Extending CDR activity based mobility patterns with census and travel survey data, Jiang et al. (2017) were able to detect specific daily mobility "motifs" of individual users and relate those to unique sociodemographic identities of different geographical areas. In another study, (Xia et al. 2018) used CDR and travel cost data (through an Application Programming Interface (API)) to estimate population flows and impedance functions of a SIM, extended to include the potential of travelling based on population density. Validated against census data, the developed model was suitable for predicting potential population flows. In another study, Chen et al. (2019) used a CDR dataset within a STP accessibility approach to estimate individual accessibility for each phone user. Evaluated at geographical cohorts, the authors found distinct accessibility patterns to shopping facilities for urban/rural and suburban groups. Further, by using taxi trajectory data, the study also demonstrated the impact of travel time uncertainties on individual accessibility. Using a combination of space-time accessibility supplemented with a cumulative measure, Chen et al. (2018) used a CDR dataset to evaluate interpersonal accessibility variation for phone users of the same residential location. In this way, the authors demonstrated the value of passive mobility data for individual based accessibility analysis.

Using a cumulative accessibility measure together with taxi GPS data, Cui

et al. (2016) detected areas of low accessibility and related the results with the levels of available activity types at a destination. Moya-Gómez et al. (2018) used car GPS data together with social media data for computation of Origin Destination (OD) times and the distribution of attraction factors. Using dynamic extension of a cumulative accessibility indicator, the authors investigated the temporal patterns of accessibility for the city of Madrid, Spain.

Smith, Quercia & Capra (2012) used a gravity model in combination with automatic fare collection (AFC) data to quantify the levels of accessibility for employment and leisure related activities. They concluded that it is likely that richer datasets are needed, in the form of socioeconomic and environmental characteristics, to derive realistic accessibility levels using pervasive mobility data. Using GTFS data, Stępniak & Goliszek (2017) derived OD matrices in the context of calculating a potential accessibility measure. In this way, the authors demonstrated the importance of considering uncertainty in diurnal fluctuations of accessibility.

2.3.2 Passive mobility data for inferring accessibility components

More granular accessibility analysis at the individual level requires derivation of semantic knowledge of attributes such as the transportation mode used and the types of activities reached. These are not readily available in passive mobility data and have to be inferred from mobility data often coupled with secondary information. This section provides a brief overview of the different computational methodologies for these tasks.

2.3.2.1 Transportation mode inference from unlabelled mobility data

Knowledge of the share of transportation modes used is very important semantic information related to an individual's mobility, and forms the basis for disaggregated evaluation of secondary mobility quantities such as number of trips per mode, access to transit etc. It is also a very volatile quantity that depends on individual characteristics such as age, gender and disability to a large extent (Ryan et al. 2015, Nordbakke 2013). The task of transportation mode detection from unlabelled passive mobility data is largely treated as a machine learning classification/clustering problem in the literature.

A popular classifier used within the transportation mode detection literature is Support Vector Machines (SVM). A SVM is a supervised linear classifier that uses a kernel function to transform the original variables into a higher dimension feature space in order to tackle the problem of linear inseparability between different categories. A common approach to tackle the problem of inseparability between travel modes, is to expand the feature space by using more quantities

e.g. using both speed and acceleration. However, such measurements might not be available in the first place. Moreover, SVM classification methods ignore the temporal structure of human mobility data, although there have been attempts to circumvent the problem (Bolbol et al. 2012). Modifying SVM models to include a wider range of information in the classification problem can be done either by altering the kernel function, or by building the model within a regression framework. As SVMs are supervised classification models, they require a training set which might not be available beforehand. As SVMs are by definition non-probabilistic classifiers, it is difficult to assess the uncertainty in the estimates over the set of classes.

Another commonly used classification approach to transportation mode detection, often thought to be one of the best performing classifiers for this task (Jahangiri & Rakha 2015) is Decision Trees (DT) (Zheng et al. 2008, McGowen & McNally 2007, Griffin & Huang 2005, Reddy et al. 2010). These can appear alone or in combination with a MNL regression model. A DT classifier recursively segments the feature space in a binary fashion, based on the principle of minimising some loss function (eg. chi-squared, entropy etc.). An elementary example of a DT classifier for determining transportation modes using speed would be to "make decisions" on the mode based on how high or low the speed value is.

DTs have the advantage of being direct and easily interpretable. However, they tend not to generalise well as they refer to a particular setting of decisions configurations only. Another big disadvantage of DTs is the large variance, especially in the case of correlated features (Janssens et al. 2006). An improvement over this, is the merging of a set of individual DT classifiers into a single one (Random Forests (RF)) to smooth the individual variances. The downside of using this method is the loss of interpretability.

A third family of models commonly used for transportation mode detection from mobility data is Artificial Neural Networks (ANN) (Zhang et al. 2015, Stenneth et al. 2011, Shafique & Hato 2015). ANNs are used to approximate complex functions by summing together weighted versions of simpler functions (neurons). These neurons can have a sigmoid response function in case of binary and categorical variables or linear response function in the case of continuous variables. For transportation mode detection, the complex function can represent the boundaries between different mode categories. Advantages of Neural Network methods include the easiness of including a wide range of variables in the classification process in a straightforward way (Omrani 2015). A disadvantage is the loss of interpretability of the classification results due to the dense network of neurons. This fact can make the generalisation of a learned ANN to datasets of different

spatial resolution difficult.

Another family of models used in transportation mode detection from passive mobility data is generative models. These use the joint probability of all the variables in the feature space together with the class probabilities to solve the classification problem. Contrary to discriminative classifiers (e.g. SVM, RF, ANN), they don't define the classification process using boundaries, but rather probability distributions that characterise the classes. This family of models include various versions of probabilistic graphical models (PGMs) such as naive Bayes, Hidden Markov Models (HMM), Bayesian Networks (BN) etc.

The simplest classifier in this context is naive Bayes. This method assumes complete independence over all variables in the feature space given the class, a condition which is difficult to defend in most cases. As a classifier, it has been found to have a reduced accuracy in the context of transportation mode inference compared to other classifiers when the feature space is limited to quantities such as speed, acceleration and heading (Reddy et al. 2010, Stenneth et al. 2011). On the other hand, in the case the feature space is broadened with variables such as distance to metro or bus lines, naive Bayes has been found to perform better than any discriminative classification method (Feng & Timmermans 2016).

BNs offer an improvement over the conditional independence assumption of naive Bayes and as a consequence, they are able to model more complex relationships between variables in the feature space. Such a model has been found to perform well in transportation mode classification by modelling the conditional relationships of acceleration, speed, trip distance and speed percentiles (Xiao et al. 2015).

The above described models tend to overlook the temporal dependence of mobility data. Dynamic models such as HMMs and Dynamic Bayesian Networks (DBN) attempt to address this issue.

HMMs are Bayesian networks exploiting the sequential nature of time stamped data. The main assumption is that an unobserved "hidden" time dependent process is the driver behind the observations. HMM are memoryless models, in the sense that a node is dependent only on the preceding node and not on the previous ones. This assumption can be relaxed if a higher order HMM is employed, In this case, however, there is a risk of over-smoothing, making classes less distinguishable. Richer modelling specification frameworks, such as a combination of HMMs with DTs, have been found to provide increased classification accuracy for different modes (Reddy et al. 2010).

DBNs combine the graph structure of Bayesian Networks with the sequential structure of Markov models. By treating problems as time dependent stochastic

processes, dynamic Bayesian networks can not only capture the associated uncertainty for each node, they can also reason about the way these evolve over time(Koller & Friedman 2009). This is due to the causal network structure of such models which allows researchers to "inject" domain knowledge in their models. Their flexibility and granularity made these models popular amongst a variety of disciplines such as speech recognition, automatic handwritten character recognition and DNA sequencing. These models are being increasingly used within the human trajectory mining and activity recognition held within an unsupervised classification framework (Liao, Patterson, Fox & Kautz 2007, Lin & Hsu 2014). A disadvantage is that, being unsupervised classification algorithms, with given additional data the parameters of the models have to be learned again (Lin & Hsu 2014).

Table 2.1 below summarises the advantages and disadvantages of some of the cited methodologies.

Table 2.1: Comparison of different transportation mode detection methodologies

Authors	**Method**	**Accuracy**	**Sensor**	**External parameters**	**Advantages**	**Limitations**
Bolbol et al. (2012)	SVM	88%	GPS	No	Includes a variety of transportation modes	Preprocessing required, inseparability issues between bus and underground mode
Janssens et al. (2006)	BN and DT	53.60%	Travel Diaries	Yes	Includes variety of information related to transportation habits	Combined method underperformed compared to BN and DT alone
Zhang et al. (2015)	ANN	62%	GPS	No	Fast computation even for large datasets	A preprocessing step is required as inference is based in a derivative of speed
Feng & Timmermans (2016)	Naïve Bayes	99.40%	GPS	Yes	Includes a variety of classification features and external parameters	Data obtained from a dedicated GPS logger with a variety of accuracy measures which are beyond the reach of low end GPS sensors

Xiao et al. (2015)	Bayes Nets	90%	GPS	No	Accounts for the inter-dependencies between classification feature space	Potential loss of information through discretisation of continuous variables
Reddy et al. (2010)	HMM	95.8%	Smart phone	No	No preprocessing step	Results were tested on one particular smart phone device only
Liao, Patterson, Fox & Kautz (2007)	DBN	75%	GPS	No	No preprocessing step, no training set	Complex model

2.3.2.2 Activity type inference from unlabelled passive mobility data

Knowledge of activity types performed at a destination is one of the most important components of accessibility studies and one of the most challenging tasks for unlabelled mobility data. Different data sources are considered for this task: examples in the literature range from AFC systems (Zhao et al. 2007), GPS traces (Shen & Stopher 2014) and CDR data, to mobility data from location enabled applications (such as Twitter and Foursquare). To assist the inference process, such data are commonly coupled with secondary information such as land use and destination data such as Points of Interest (POI) and land use data that could inform on the nature of the performed activities.

In nearly all cases of human activity inference from mobility data, the range of activities is commonly discretized to a finite set such as home, work, leisure, shopping etc. Methodological approaches on the task of inference on this discrete set of activities varies depending on the nature of the data as well as the research goal. For example, when the primary focus is the generation of activity data for the purposes of transportation demand modelling (eg. origin destination matrices), an activity type set consisting of home/work locations is sufficient. Accessibility modelling however, relies on an expanded set of activities that span beyond commuting to/from work such as recreation and shopping. Relying on spatiotemporal characteristics of mobility data is not enough, as such activities often fall beyond any regularities that can be leveraged from such data. As a result, researchers have looked into other, complementary sources of information that can compensate for this limitation. The most straightforward approach that has been proposed in the literature is to associate each transportation access point with an activity that is a function of the POIs that exist within a predefined distance from the transportation access point (Chapleau et al. 2008) . Different methodological frameworks made use of such data ranging from simply counting the number of POIs within a distance from a public transport stop/station (Long & Shen 2015), to more complicated ones that attempt to address the fact that POI do not necessarily correlate with the actual activities that are undertaken at a particular location (Hasan & Ukkusuri 2017). In practise, however, a combination of different approaches is applied to assist inference. In terms of output, different approaches produce different outputs ranging from activity type clusters using similar characteristics of the input feature vector to probabilities of specific activity types. Usually the former requires an extra interpretation step to derive semantic context from the clusters.

Rule based activity detection methods are one of the most used methods of imputing activities from mobility data. Such methods have been successfully used

in the context of transport data for the determination of activities such as home, employment and study (Anda et al. 2017). The rules generally follow from assumptions related to temporal regularities of different activities, together with assumptions on the travel frequency as well as the spatial distance between subsequent destination locations as derived from mobility data (Barry et al. 2009). Commonly, classification rules are derived from past travel survey data (Wang et al. 2017), however it is not uncommon to derive such rules from behavioural patterns in the mobility data, especially when the activity space set consists of predictable categories such as employment (Lee & Hickman 2014).

Example applications of rule based methods can be found throughout the literature for activities such as home, employment and a general category that captures the remaining activities (such as shopping, entertainment etc.). Using AFC data, (Long & Shen 2015) and (Wang et al. 2017) defined a "home" activity station to be the station where the first trip of the day is made, as long as this pattern is consistent throughout the sequence of AFC observations. To determine an individual's workplace station, the authors added a temporal threshold to the remaining stations along a user's daily AFC observations not categorised as home. In the context of these studies, this threshold was determined from past origin destination surveys. Devillaine et al. (2012) approach was also based on a set of empirical rules derived by querying the AFC records database (for the cities of Santiago and Gatineau) to distinguish home, work, study and "other" activities. They did this by specifying hard temporal thresholds together with the record's sequence in relation to the time of transaction. The authors discovered similar activity patterns between these cities despite their unique socioeconomic characteristics.

Moving away from AFC data, Alexander et al. (2015) used call detail records from mobile phones to infer important places such as home and work. Due to the noisy nature and the reduced spatial resolution of the data, the authors had to agglomerate the individual location estimates into clusters of location data, before extracting features such as duration of stay. Spatio-temporal rules were then applied to those features to distinguish home and work locations. Specifically, the authors defined a temporal window within which an individual is expected to be home, and a spatiotemporal window for work location that combined the observations falling into a temporal window on weekdays along with a spatial distance threshold reflecting the assumption that longer distance trips are more likely to be work trips (Levinson & Kumar 1994).

One disadvantage of the above reviewed activity inference methods is the reduced flexibility to model more complex relationships between activity types, attributes derived from mobility data as well as secondary information such as

characteristics of the built environment. Moreover, quantifying uncertainties originating from the noisy, inaccurate and incomplete nature of mobility data using rule based methods and heuristics is difficult. Probabilistic methods can account for this either by representing such relationships through a set of conditional probabilities between the latent activities and the feature space variables (discriminative models) or modelling the joint distribution of activities and feature space variables (generative models). This relationship is commonly represented using a graph structure that factorises the joint probability density density over the set of random variables depending on how these variables are assumed to interact with each other.

In terms of applications using generative probabilistic models, Yuan et al. (2012) used a combination of GPS and POIs to infer functional regions corresponding to different activity types in the city of Beijing. Following the analogy of using GPS traces as words and POIs as documents, the authors used a topic modelling framework (Latent Dirichlet Allocation, LDA) to discover regions of similar semantic background. A LDA is a directed probabilistic graphical model that uses a "bag of words" assumption to represent documents as a mixture of topics, each one characterised by a distribution of words belonging to a certain topic (Blei et al. 2003). Within a similar modelling framework Hasan & Ukkusuri (2014) used the analogy between check-ins/words activities/topics to geo-tagged Twitter feeds linked to Foursquare check-in data. Their model was able to classify individual check-ins into higher level activities such as entertainment, education and shopping, however, it is unclear how the above approach can be applied to data without any semantic reference such as un-labeled social media data or mobility data of comparable granularity. Furthermore, it is unclear how LDA operates in the context of sparse feature vector scenarios such as limited observations (few documents), and short observation vectors (documents with few words). Both of these cases are characteristic of mobility data generated by service providers such as Automatic Fare Collection systems where there are as few as two interactions of an individual with the transportation system per day, or in scenarios where the POI feature vector is limited to few POIs (as in the case of less dense urban environments).

In another study, Widhalm et al. (2015) used the concept of Relational Markov Networks (RMNs) to impute activities such as home, work, shop and leisure from functional clusters derived from cell tower mobile phone and landuse data. RMNs are an extension of Markov Random Fields, modelling the factor potentials in a structure that resembles a relational database. Specifically, their model specified the activities given land use types, activity duration, starting time as well

as heuristic rules (eg. if the activity was visited previously, if the activity has a unique location). Their approach achieved comparable results in the activity clusters compared to traditional origin destination surveys, however, here an extra interpretation step is needed to extract specific activity types from discovered clusters.

Table 2.3 below, summarises the advantages and disadvantages of different activity inference methodologies found in the literature.

Authors	Method	Validation	Scope	Advantages	Limitations
Alexander et al. (2015)	Rule based (frequency of visits, distance, time)	Using survey data	CDR / identifying home, work and 'other'	Can be applied to large datasets	Approximate validation / limited activity types
Long & Thill (2015), Wang et al. (2017)	Rule based (trip sequence, location , time)	Using survey data	AFC / identifying home, work	Can be applied to large datasets	Approximate validation / limited activity types
Yuan et al. (2012)	Latent Dirichlet Allocation (LDA)	Performance bench marking using different algorithms	GPS / Functional clusters	Inclusion of both mobility and POI data	No absolute validation / requires an annotation step
Hasan & Ukkusuri (2017)	Continuous time bayesian network (CTBN)	Using synthetic data	Twitter check-ins / Functional clusters	Good accuracy on activity transitions	Approximate validation / No semantic information on activities
Han & Sohn (2016)	Hidden Markov Model (HMM)	Using survey data	AFC / Functional clusters	Accurate representation of activity sequences	Approximate validation / No semantic information on activities
Yin et al. (2018)	Input-Output Hidden Markov Model (IOHMM)	Using survey data / small sample of ground truth	CDR / Identifying home, work, eating, recreation etc.	Includes secondary activities	Activity types are determined empirically
Xiao et al. (2016)	Artificial Neural Network (ANN)	Participant validated	Smartphone data / Identifying home, work, eating, shopping etc.	Includes a wide range of features	Requires labelled mobility data
Liao et al. (2006)	Conditional Random Field (CRF)	Participant validated	GPS / Identifying home, work, leisure, visiting	Good activity detection accuracy	Very small sample / Requires labelled mobility data
Widhalm et al. (2015)	Relational Markov Network (RMN)	Using survey data	CDR / Functional clusters	Database relationships can be directly mapped to factor potentials	Approximate validation / requires an annotation step
Bantis & Haworth (2019)	Dirichlet Multinomial Dynamic Bayesian Network (DBN)	Participant validated	Foursquare check-ins / Food, Shopping, Outdoors and recreation, Arts and Entertainment, Colleges and Universities	Accuracy in par with state of the art, no annotation required, allows incorporation of different levels of prior belief	Computationally intensive / not suitable for real time applications

Table 2.3: Comparison of different activity inference type methodologies

2.4 Accessibility within social sciences

Having introduced the most widely used accessibility indicators and the challenges and potentials of using passive mobility data in accessibility evaluations, this section introduces the necessary research background on the link between accessibility and transport related social exclusion. The goals of this section are to highlight the different dimensions that contribute to the process of transport related social exclusion, as well as to provide insights on how accessibility indicators are used to study issues of social exclusion and equity.

2.4.1 Defining accessibility in a social sciences context

When the research focus is the connection between accessibility and social processes that can result in relative disadvantage, such as transport related social exclusion, the term accessibility is generally viewed as a fundamental property of individuals' ability to participate in different activities within civil society (Burns 1980, Preston & Rajé 2007). The definition refers to the extent to which a person is able to reach a range of destinations that can facilitate the completion of different social, leisure and employment activities considered to be normal in a particular society (Evans 2009, Nutley 1998). This ability takes the wider urban environment characteristics into consideration, such as transport provision (buses, trains etc.), environmental characteristics as well as individual preferences and capabilities (Farrington 2007, Kwan 2013). Related to this, Church et al. (2000) identified seven distinct factors that could reduce access to opportunities, which are shown in table 2.4.

Category	Description
Physical Exclusion	Physical conditions can affect the ability of the person to effectively use the transport network. Examples are people with reduced mobility, learning disabilities, visual impairments or age related difficulties
Geographical exclusion	Longer commutes from place of living and workplaces or amenities can make them difficult to reach due to temporal or financial aspects
Exclusion from facilities	Lack of transport services within the area is likely to cause transport exclusion, particularly for people who do not own a car
Economic exclusion	Whenever low financial income affects the ability of a person to afford transport costs
Time based exclusion	When personal factors or duties reduce the time available for travelling. This phenomenon is particularly common among carers who lacks of an adequate social support network
Fear-based exclusion	Fear for personal safety can easily discourage individuals from using the transport network or other public spaces
Space exclusion	Space management can affect the perception and consequentially the use of public spaces and services.

Table 2.4: Factors related to social exclusion as identified by Church 2000 (p.198-200)

It is important to note that the factors in table 2.4 tend not to appear in isolation, and coexisting factors are more likely to increase the risk of transport related social exclusion.

Although the above description of accessibility overlaps with the notion of mobility, it also highlights some key concepts that tend to be overlooked by thinking only in terms of mobility. Traditionally in transportation planning and engineering, individual mobility refers to the resources and characteristics of individuals (financial status, age, access to a car etc.) that facilitate the ability of a person to move from place to place (Tyler 2006). However, increased mobility does not necessarily result in increased accessibility. For example, a person can be thoroughly mobile and still experience barriers when attempting to reach an activity. Besides

physical and geographical, these barriers could be of a social nature as shown in table 2.4. This highlights the need to consider factors related to both mobility (ability to use different transportation modes) and accessibility (ability to reach opportunities) in the modelling approach.

2.4.2 Accessibility equity and transport related social exclusion

The link between transport disadvantage and issues such as social exclusion, well-being and discussions around issues of equity and equality has been recognised since the 1960's. Fairly recently, this discussion was extended to recognise the fundamental role of accessibility on such issues (Pereira et al. 2017, Lucas 2012, Casas 2007). According to a widely cited definition by Kenyon (2003), transport related social exclusion is a process by which individuals are prevented from participating in different aspects of a social life in a community. This may be because of reduced access to opportunities, services and social networks or due to insufficient mobility in a society. Such a process leads to decreased levels of well-being, particularly for vulnerable population groups (Currie et al. 2010).

Another ethically relevant topic is that of equity in accessibility evaluations. The focus here is on the distributional effects, or the 'fairness' of access to opportunities (Van Wee & Geurs 2011). The relation with transport related social exclusion can be summarised by the premise that equity dictates the level of potential to participate in opportunities, regardless of factors such as age, gender, ethnicity, race or income, to name just a few. Lower levels of opportunities due to lack of transport results in transport related social exclusion. Judgement of redistribution (or what is considered fair) of accessibility is a complex topic as it requires a degree of moral judgement and as such it is often unaddressed in equity related accessibility evaluations (Farrington & Farrington 2005, Van Wee & Geurs 2011). Current guidance on methods and processes to study equity are not comprehensive enough and as a result, stakeholders (transportation agencies, public bodies etc.) wishing to evaluate equity often resort to simple accessibility indicators. These are not always insightful to identify relative disadvantages (Karner & Niemeier 2013). In a review of accessibility implementation plans conducted by the relevant associations/institutions of 14 North American cities, (Manaugh et al. 2015) found that, beyond recognising the existence of accessibility related issues, there is little evidence of specific measures that can address challenges such as transport disadvantage, social equity and quality of life. As a result, the majority of measures focus on challenges that can be tackled in a more direct way. Examples include environmental aspects such as reducing carbon emissions as well as mobility related issues such as optimising proximity to available transportation

access points. To that extent, framing accessibility evaluations around theories of social justice can help provide the basic framework within which such evaluations of equity can be made (Pereira et al. 2017).

A social exclusion approach to transport disadvantage puts the focus on the outcomes of transport deprivation (Titheridge et al. 2014), however, it is important to notice that this concept emphasizes both the causal factors that lead to such a condition and the interactions between them (Lucas 2012). As already mentioned in section 2.4.1, these factors include characteristics that lie with the individual, characteristics of the local area as well as wider economic societal and governance factors. The lack of available transport options or inability to use them, together with disadvantaged personal status, reduces the ability of an individual to reach different opportunities. This causes a lack of accessibility, which in turn is manifested as social exclusion. The causal flow between transport disadvantage, social disadvantage and social exclusion as well as the factors that influence them is illustrated in figure 2.4 (Lucas 2012):

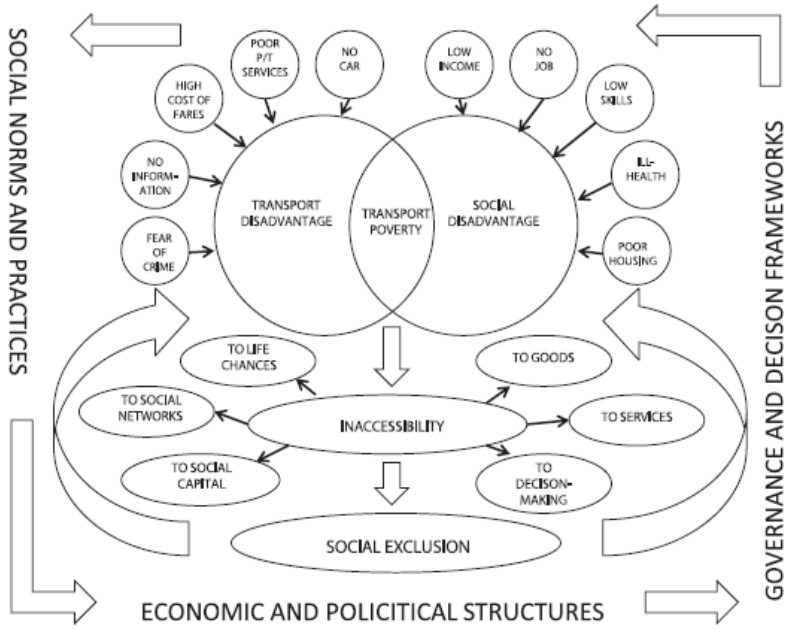


Figure 2.4: Causal structure between transport disadvantage, social disadvantage accessibility and social exclusion as illustrated by (Lucas 2012)

In the same line of thought, Preston & Rajé (2007) argue that the effects of social exclusion are not due to lack of social opportunities, but because of lack of access to those opportunities. According to the authors, addressing social exclusion requires extending the knowledge of person/place interaction beyond transport geography and into the domain of social-spatial research. Approaching

accessibility from this angle, Farrington & Farrington (2005) redefined the terms used to describe accessibility. Opportunities become more than locations on a map; they are potentials for achieving one's needs, wants, aspirations and desires. Reaching opportunities becomes more than a property of space or a property of the transport system, as it doesn't necessarily reflect the ability of an individual to participate in the activities associated with each destination (Pereira et al. 2017).

At this point, it should be noted that case studies seeking to quantify transport disadvantage or transport related social exclusion[4] rarely adopt the above described definition of accessibility in its entirety. Instead, existing accessibility indicators covering aspects of the above definition are used (Kamruzzaman et al. 2016, Pyrialakou et al. 2016). On the other hand, when the focus of the studies is the use of accessibility indicators to examine the link between accessibility and equity, engagement with theories of social justice is rarely made (Pereira et al. 2017)). Finally, it should be mentioned that accessibility is only one way of quantifying transport related social exclusion, albeit the most holistic one. Other methods include structured questionnaires and basic statistical analysis (Delbosc & Currie 2011*a*), outcome based analysis such as measurement of individual activity spaces (Schönfelder & Axhausen 2003), deprivation based measures (Noble et al. 2007), mobility based measures (Dodson et al. 2006) and structural equation models (Golob & McNally 1997).

Examples of how the most commonly used accessibility indicators are used with transport equity in the literature are given in the following sections.

2.4.2.1 Gravity-based accessibility indicators

Preston & Rajé (2007) used three criteria to evaluate the social exclusion process: levels of travel in an area, levels of individual travel and the overall accessibility of an area (as assessed by gravity and utility based accessibility indicators. The authors raised issues of data availability (in particular extensive surveys of individual travelers) as barriers for adopting a more disaggregated approach, both spatially and socially, recognising that current accessibility planning tools are not sensitive to issues such as gender, age and ethnicity. Generation of synthetic population mobility data as an alternative to data availability is unlikely to provide a solution as the major weakness of such methods is identifying the unique combinations of attributes of individuals affected by social exclusion (Preston & Rajé 2007). Bocarejo S & Oviedo H (2012) used a gravity based indicator to capture the effects of income and travel time for travelling to employment. This was done by modifying

[4]Although transport disadvantage and transport related social exclusion are different concepts, the indicators used in case studies are often identical (Kamruzzaman et al. 2016)

the impedance function of the model to include travel cost and travel time in the context of evaluating investments in transport provision that promote equity in the city of Bogota. The authors disaggregated the results geographically and by income strata. However, both the sets of accessibility components used and activities reached were restricted. Again, one of the main reasons for this was data availability. Recently, transit data in GTFS (General Transit Feed Specification) format have been used by researchers to obtain disaggregated transport data. For example, Karner (2018) used GTFS data with OD matrices to assess transport equity for different income groups in Phoenix, USA. In their analysis, they used a disaggregated gravity type model for each wage level. Although useful, GTFS data are essentially timetables which describe the range of transport options available and not their usage.

2.4.2.2 Cumulative-based accessibility indicators

Wu & Hine (2003) used a contour based accessibility approach (Public Transport Accessibility Levels (PTAL)) in combination with deprivation indices to identify issues of transport disadvantage in households living in areas with limited transport coverage and investigate how these would change under different infrastructure change scenarios in the city of Belfast. The study captured structural differences in transport provision for different religions and age groups. However the analysis was conducted at a level of aggregation that prevents a more detailed study of individual characteristics in relation to mobility and accessibility patterns. In another study (Ben-Elia & Benenson 2019) used a combined cumulative accessibility index composed of the total travel time by public transport and car and the total count of destinations at building level to study to study issues of spatial equity in the city of Tel Aviv. The authors focused on providing a disaggregated approach to evaluating equity using accessibility. However, the study did not account for sociodemographic characteristics of individuals that may influence the composite accessibility index. The GTFS format was also explored to obtain more detailed data for this accessibility indicator. For example, (Farber & Fu 2017) used a cumulative accessibility indicator to explore how accessibility is changed under different network modifications.

2.4.2.3 Space-time accessibility indicators

Using a space-time accessibility indicator, Fransen & Farber (2019) evaluated the levels of equity experienced by individuals in the wider area of Utah, USA. They found that both place based and sociodemographic characteristics resulted in significant differences in accessibility levels. Kwan (1999) used the concept of poten-

tial path area of space-time accessibility to study the significance of an individual's ethnic background and gender when accessing day-to-day destinations. The authors found that the levels of individual access in women is significantly lower compared to those of men. In a study comparing different methods of quantifying social exclusion of children from available opportunities, Casas et al. (2009) found that space-time accessibility measures can adequately represent children's activity spaces given household socio-demographic characteristics and carers' time budgets. However, a common modelling trait of studies using space-time accessibility measures is that they don't explicitly include individual characteristics in the modelling process and as a result, accessibility outputs have to be related to different population cohorts as a separate step in the analysis. This requires detailed travel diary and time use data and as a result, the number of participants is limited and the geographic extent of the studies relatively small (Geurs & Ritsema van Eck 2001, Neutens et al. 2011). Fairly recently however, researchers have begun to explore the usefulness of machine generated data to obtain travel time distributions at the individual level. For example, Chen et al. (2019) used mobile phone cell tower data with a space-time accessibility approach to estimate travel times, and evaluate all feasible space-time locations to perform a flexible activity. They then combined the results with a cumulative accessibility indicator to detect equity issues in accessibility to shopping facilities for different geographical groups (urban/rural/suburban). Using GTFS data to model the state of the transport system before and after the introduction of a new bus service in Colombus Ohio USA, (Lee & Miller 2018) used a space-time accessibility measure together with a cumulative opportunities measure to investigate the impact of this new service on access to job and healthcare facilities. The authors were able to identify the optimal combinations of services that would result in increased accessibility. However, the study didn't take into consideration any behavioural factors that would result in different accessibility levels, rendering the approach mostly relevant as a physical infrastructure accessibility indicator.

2.4.2.4 Utility based accessibility indicators

Comparing the performance of different accessibility measures on the equity of access of different destinations for the city of Ghent in Belgium, Neutens, Schwanen, Witlox & De Maeyer (2010) found that utility-based measures better articulated interpersonal differences, providing more conservative estimates of equity. Using a utility based accessibility indicator (the logsum measure) evaluated on travel demand survey data, Bills & Walker (2017) compared changes across and among low and high income population groups with regards to equity in transportation

provision under different travel cost and time scenarios. The authors generated useful insights on the shape of transportation mode share under those scenarios. However, the study makes a major assumption that by using the observed mode to calculate the equity indicator, the individual acts as a maximiser. This is restricting in the context of representing the potential behaviour of an individual as observed behaviour might not necessarily reflect desired behaviour. In the context of assessing the sensitivity of different equity measures for different transportation provision policy scenarios, Ramjerdi (2006) used the logsum indicator to compare levels of equity across different regions of Oslo, Norway. The authors found that equity levels were sensitive to different spatial aggregation levels of the indicator. At this point, it should be noted that when the focus is economic evaluations of accessibility, utility base indicators outperform all other accessibility indicators (Geurs 2018, Bhandari et al. 2009).

2.5 Accessibility through the lens of social justice theories

The discussion of section 2.4.2 highlighted the usefulness of accessibility indicators for the task of evaluating issues of transport related social exclusion and equity experienced by individuals and/or population groups. Studying the indicators can be useful in themselves (Van Wee 2016). However, a closer engagement with social justice theories and the insights they generate when applied to transport equity could provide a transparent framework both for analysing (by providing structure and identifying important factors and their interactions to be included in the analysis) and assessing (by providing a direction on the measurable quantities that could be used to study the process of transport related social exclusion) issues of equity in transport (Mullen et al. 2014).

2.5.1 Accessibility indicators from a social justice perspective

From an equality perspective, the link between accessibility, and transport related social exclusion and equity has been examined through the lens of different theories of social justice such as utilitarianism, libertarianism, sufficientarism, Rawl's egalitarianism and the Capabilities Approach (CA) (Pereira et al. 2017, Lucas et al. 2016). In this context, social justice in transport refers to the fairness of distribution of goods, transport services and accessibility for people (Beyazit 2011).

Approaching the topic through these different schools of thought results in different interpretations of the wider definition of accessibility. For example utilitarianism primarily focuses on the instrumental value of travelling to activity

locations in order to benefit from the activities that take place at those locations. As such, it is not a goal in itself, but rather a tool that facilitates utility maximisation. This approach has been adopted by transportation providers primarily through willingness-to-pay surveys for different transport demand models (McFadden 1998). However, this approach has been criticised on the grounds that it usually focuses on monetising values such as travel time, distance and convenience of services and as such it implicitly emphasises activities that provide higher value, such as employment (Van Wee & Roeser 2013). Moreover, since utilitarians seek to maximise the benefits (utility) for the whole society, they tend to approach accessibility on an aggregate level, not paying attention to how it is distributed throughout society (Pereira et al. 2017). This leads to under-representation of the most vulnerable individuals within a society in accessibility evaluations, or even worse, the promotion of accessibility for people who are better off at the expense of those who are not.

The idea of libertarianism dictates that people should be able to keep what they earn or inherit (Titheridge et al. 2014), with minimal intervention by the state or others, provided that the rights of the rest of the people to do the same are respected. Within an accessibility discussion, the accessibility benefits from an intervention are distributed in accordance to the rules that dictate a free market. Currently, there is a major push to support and expand initiatives related to "Mobility as a Service" (MaaS) from both research institutions and government funds (UKGovenrment 2018). Many applications of MaaS fall under the economies of sharing applications such as Uber and Lyft. In this context, Pereira et al. (2017) mentions that the economy of sharing services may have expanded the choices of transportation of consumers, however, it raises issues of fairness and equity among conflicting sectors and among minority population groups. This is because private companies have no economic benefit from attending to the special needs of such groups.

Sufficientarianism theories assume that people should be 'well-off' up until a certain minimum threshold deemed sufficient to guarantee that basic needs and wellbeing is maintained (Lucas et al. 2016). The theory provides a justification for relative agencies to set a minimum threshold of accessibility below which individuals falling below are considered socially excluded. Given the complexity of the processes leading to social exclusion and the diversity of people's needs, a big question here is what constitutes a minimum threshold of accessibility and who decides what it is. This question remains unanswered in the relevant literature (Preston & Rajé 2007) . Expressing accessibility indicators in the context of determining a minimum threshold is generally related to the principles of sufficientarianism. Ini-

tiatives such as that undertaken by the Social Exclusion Unit in the UK are a step in this direction. However, as Farrington & Farrington (2005) puts it, given the fragmented nature of assets and responsibilities of different stakeholders (examples can be: land use planning departments, health institutions, education institutions, transport authorities, social services, citizens etc.) that could play a role in improving accessibility, it is unclear who is responsible for setting thresholds that cut across different organisations. It is argued that it is unlikely that an overarching organisation (for example local authorities as suggested by Social Exclusion Unit) hold enough political influence to facilitate cross sector collaboration between the different players.

Rawl's egalitarian approach is based around the premise that all people should be treated equally and have as much freedom as possible, as long as this doesn't compromise the freedoms of others. This premise is extended to allow for the difference principle (Martens & Golub 2012), meaning that inequalities in the distribution of primary goods should be allowed to exist, as long as they benefit the least advantaged members of society. This reading promotes the role of agencies that can facilitate equality in distribution of goods, such as institutions (Pereira et al. 2017). Many authors argue that accessibility in its wider definition should be counted as a primary good (Van Wee & Geurs 2011, Khisty 1996) and, through Rawl's egalitarian approach, it is the role of institutions and policy makers to ensure that any interventions in transport aimed at improving accessibility, will do so for the least advantaged groups. One point of critique of the difference principle in Rawl's egalitarian approach is that it doesn't differentiate between inequalities resulting from arbitrary circumstances (such as being born poor) and those resulting from personal choices. As a result, any interventions aimed at mitigating arbitrary circumstances will result in mitigating the legitimate choices of others (Pereira et al. 2017). Current accessibility indicators that focus on the differences between people generally agree with the equality principles of egalitarianism. Indeed, the three step process of selecting appropriate accessibility indicators, calculating accessibility for different population groups, and comparing the changes in the indicators across groups and across different scenarios followed by most studies (Bills & Walker 2017) generally reflect those principles. However, the specifics of accessibility indicators can produce very different results, even if they fall under the egalitarian philosophical theory. Martens & Golub (2012) have distinguished three different approaches to equity: equality of resources, equality of midfare and equality of welfare. Simple infrastructure reliability measures as well as simpler cumulative accessibility indicators fall under this category. Focusing on actual accessibility patterns, utility based and space-time based indicators that focus on

actual behaviour fall under the welfare approach. More complicated cumulative and gravity type indicators can be categorised in the equality of midfare approach.

The CA was first introduced by the philosopher and economist Amartya Sen in the 1980's (Sen et al. 1990), and was originally developed as an alternative to the predominant utilitarian way of viewing notions such as quality of life and well-being in welfare economics. Its success as a theory of social justice has led to the creation of the Human Development Index[5] by the United Nations Development Programme for the purposes of ranking countries by their level of well-being. In essence, the CA describes the ability of an individual to function given the set of freedoms and practical opportunities that are available to them (Sen et al. 1990). Contrary to Rawl's egalitarian approach, where the emphasis is on primary goods (Rawls 2009), the CA focuses on human capabilities which result from a combination of personal abilities, and the wider environment (Pereira et al. 2017). In this sense, the CA has many similarities with the egalitarian midfare approach since it focuses on the ability of an individual to convert resources into welfare, a compromise in essence between resources and welfare (Martens & Golub 2012).

The CA can be perceived as a normative evaluation concept, aimed at promoting public policies towards improvement of the abilities of individuals to function as opposed to just describing the problem. This allows for the relative assessment of different policy proposals and the effect that those will have on a person's well-being (Alkire 2008). As accessibility has been traditionally been used as a tool that can push towards policy changes (Pirie 1981) the CA seems to fit within that framework. Viewing accessibility within this context encompasses not only the ability of individuals to move so that they can conduct the activities they value, but also includes all the policies that enable people to do so (Pereira et al. 2017).

However, the application of the CA within an accessibility framework is challenging for a number of different reasons. First, it doesn't prescribe thresholds on the minimum accessibility needs of individuals. This limits initiatives to improve accessibility levels of people that fall below such thresholds (Hananel & Berechman 2016). Another issue that is commonly brought in the agenda is that of individualism. The argument here is that the CA framework fails to account for capabilities which arise from collective participation since it is focused on the individual level. As such, there is the risk of omitting in the analysis the institutions, public bodies and communities that helped create and sustain the capabilities in the first place (Deneulin 2008). However, some researchers (Alkire 2008) argue that such intrinsic importance to group capabilities should at least be viewed with caution since participation in collective activities on its own is not enough to evaluate

[5] http://hdr.undp.org/en/content/human-development-index-hdi

well-being on the individual level. As an example, the use of specialised public transport services such as London's Dial-A-Ride could be claimed to be a good way of expanding the capability of "being mobile" for the group of people using it. However, if a person doesn't enjoy using this service (because of the reduced flexibility offered, having to go through the booking procedure etc.) but is obliged to use it because no other travelling option is available then that person does not "trade" that capability with increased well-being. Moreover, on a practical level, bringing this framework to real-life applications implies that a robust methodology exists that derives accessibility as a capability, taking into consideration personal abilities, socio-economic factors and the built environment. Due to this complexity, this notion of accessibility is rarely used in transport studies (Tyler 2006). A more detailed description of the CA in the context of accessibility is given in chapter 3.

2.6 Chapter summary

This chapter provided the research background of this thesis by introducing the most common accessibility indicators and briefly describing the different components influencing their relation to equity in transport and transport related social exclusion. The placement of accessibility indicators within theories of social justice was given along with a comparison table across different considerations from both technical implementation and equity perspectives. At the same time it was postulated that the CA could provide a platform to express issues of social justice while accounting for the complexity of interactions between the different accessibility components. The following chapter introduces the CA in more detail and how it has been used within the accessibility literature.

Considering the practical and theoretical strengths and weaknesses of accessibility indicators as described in section 2.2, one can conclude that there is no single numerical accessibility framework encompassing all different aspects of people's access to opportunities and the implications this might have on issues of social justice and policy making. Given the different geographical scales of measurement, measurement domain and target group, different accessibility indicators can be used to address different components of social equity. However, the choice of a measure can produce drastically different results. Using the Gini index as a comparison framework, Neutens, Schwanen, Witlox & De Maeyer (2010) have found that both cumulative accessibility and gravity type models can greatly overestimate equity of access to different urban services. On the other hand, both space-time and utility indicators produce a far more conservative result. Moreover, there are con-

siderable differences within the same category of measures. This is especially true for indicators focusing on the individual. To this extent, the choice and operationalisation of accessibility indicators remains an open challenge (Geurs 2018). Consolidating the findings of this chapter, a comparison table between different accessibility indicators with respect to applications, theoretical basis, technical considerations and equity considerations is given.

	Cumulative indicators	**Gravity type indicators**	**Space-time indicators**	**Utility based indicators**
Theoretical basis	Lacking individual behaviour mechanism; generally weak	Spatial interaction modelling; GLM regression; entropy maximisation;	Time geography; Computational geometry	Welfare economics; microeconomics; discrete choice theory
Technical considerations	Can be integrated with other indicators easily; Arbitrary thresholds of contours; Invariant across individuals	Non-linear functional form; Lack of causal structure; Can capture accessibility interactions Calibration process of balancing factors greatly affects output; Propensity for overestimating accessibility for origins with big mass; Calibration parameters invariant across individuals	Calculation of realistic extent of PPA challenging; Interactions between accessibility components geometrically defined; Inherently includes temporal component	Linear additive form of accessibility components in the utility function restrictive for complex behaviours; Addressing correlations challenging for complex interactions; Inclusion of dynamic component difficult; Estimated coefficients invariant across people; Suitable for microsimulation

Equity considerations	Can be related to both egalitarian (from a potential accessibility perspective) and sufficientarianism; Related to the potential of activity participation; Increased supply of opportunities does not mean more choices; Only applicable at an aggregated level	Can be related to both egalitarian (from a potential accessibility perspective) and sufficientarianism; Related to the potential of activity participation; Accounts for competition effects; Only applicable at an aggregated level	Can be related to egalitarian principles both from potential and actual accessibility perspectives; Not suitable for competition effects; Applicable at the level of individual; Does not prescribe behavioural mechanisms	Can be related to egalitarian welfare approach and utilitarian principles; Not suitable for potential accessibility; Suitable for describing individual behaviour
Practical considerations	Easy to compute and communicate; Enables comparisons across geographic domains (e.g. cities); Modest data requirements	Calibration can be difficult for highly disaggregated models; Modest data requirements	Difficult to visualise and communicate; Results not easily generalised to population level; Detailed travel data are needed (e.g. travel diaries, time use studies, cell-tower data)	Difficult to communicate to non-experts; Case study results not easily comparable; Requires detailed travel diary data (e.g. travel diaries)

Applications	Applied in a wide range of equity studies (e.g. urban services (Kelobonye et al. 2020), healthcare (Neutens 2015), employment (El-Geneidy, Buliung, Diab, van Lierop, Langlois & Legrain 2016), infrastructure investment evaluations (Pereira 2019), food desserts (Farber et al. 2014))	Applied in a wide range of equity studies (e.g. public transport (Karner 2018), healthcare (Lowe & Sen 1996), employment (Merlin & Hu 2017)	Applied in equity studies at sub-population level (e.g. ethnic background and gender (Kwan 1999), children (Casas et al. 2009), urban-rural (Chen et al. 2019) ; Transportation infrastructure optimization (Tong et al. 2015); Urban services optimisation (Neutens, Schwanen, Witlox & De Maeyer 2010)	Applied in equity studies focusing on interpersonal differences (Neutens, Schwanen, Witlox & De Maeyer 2010), transportation provision (Bills & Walker 2017, Ramjerdi 2006); Suitable for economic accessibility appraisals (Bhandari et al. 2009); Suitable for policy evaluation under different scenarios

Table 2.6: Comparison of different accessibility indicators

Taking into consideration the arguments presented in table 2.6 some authors argue that there is a need to gain a more complete understanding of accessibility in terms of the social justice challenges they are supposed to address (Pereira et al. 2017). Approaching accessibility modelling through the lens of CA can be useful in this regard. Despite the disadvantages mentioned in section 2.5.1, the CA appears to be a promising candidate for expressing issues of social justice within an accessibility framework (Nahmias-Biran et al. 2017, Hananel & Berechman 2016, Pereira et al. 2017). This potential is expressed by two qualities of the CA: Providing a structure that describes the ability of an individual to transform available goods and services into capabilities and welfare and; prescribing the measurable quantities (capabilities) that can be used in equity evaluations.

Chapter 3

Capabilities approach and accessibility

3.1 Chapter overview

This chapter aims to introduce the Capabilities Approach (CA) as a framework to structure accessibility modelling for equity related evaluations. It begins by providing an overview of the basic concepts and premises, followed by a discussion on how the CA has been used within the transportation literature from two viewpoints: using the CA to define accessibility and using the CA framework to examine issues of equity in transport. It then argues that existing accessibility measures fall behind in capturing the basic components of the CA, and concludes by proposing a graph theoretic approach as an alternative. Some theoretical aspects of this chapter have been published in Journal of Transport Geography vol. 84 "Assessing transport related social exclusion using a capabilities approach to accessibility framework: A dynamic Bayesian network approach" (Bantis & Haworth 2020).

3.2 The Capabilities Approach

As already mentioned in section 2.5, the CA describes the ability of an individual to function given the set of practical opportunities that are available to them (Sen et al. 1990). Two notions are central in this theory:

- **Capabilities:** These refer to the practical opportunities available and are the combination of beings and doings that a person can achieve (e.g. being socially active; doing recreational and leisure activities etc.).

- **Functionings:** These refer to the various things a person may value doing and being (Sen 2014), representing what an individual actually achieves (e.g. working; shopping; taking the bus etc.).

In accessibility terms, functioning can be understood as the realisation of day-to-day activities (e.g. shopping, getting to work, visiting friends etc.). The practical opportunities constitute the capabilities that each person has to complete these activities. Although the capability set is not directly observable, it can be derived from a set of functioning vectors from which the person has the freedom to choose (Mitra 2006). In this reading, the CA can be used to capture elements of social freedom (the ability to achieve various functions and realise one's potential), the ability to convert available resources (e.g. income, education levels etc.), welfare (the ability to achieve these functions) and equity (by evaluating the 'fairness' of achieving those functions) (Hananel & Berechman 2016).

Within the CA, the notion of functioning vectors refer to all factors that shape the capabilities set. The scope of functioning vectors can be very broad and can be made to include different elements such as an individual's characteristics (e.g. age, income, impairment etc.), characteristics of the environment (e.g. social, physical, cultural etc.) or commodities (e.g. possession of a car, availability of public transport means etc.). Disaggregating the above, one can distinguish three main concepts of the CA (Lelli 2008): The first one is the functioning vectors an individual has at his/her disposal. The second is the ability to convert the elements in these vectors in a way that will result in realised functionings (such as using the available public transport). The third refers to the notion of capabilities, which can be viewed as the set of all functionings a person could choose given his/her ability to convert the elements of the functioning vector to realised functionings.

Besides the above, another concept that plays a central role in Sen's theory is agency. In this framework, an agent is someone who acts and facilitates change. This change is evaluated in terms of the subject's own values or opinions of what is important. It is important to notice that although well-being and agency are related concepts for Sen, they are not equivalent. One example commonly used is that of practising fasting. Practising fasting can result in malnutrition and as a result a degradation of one's well-being. Assuming the person has the capability to choose whether to practice fasting (for religious reasons perhaps) or eat, a choice of the former constitutes an act of agency.

This interpretation of agency is different from that encountered in the principal-agent problem, where an agent is someone who acts on behalf of someone else. However, these two definitions need not be exclusive. For example, in the case of people with mobility impairments relying on a carer to complete their day-to-day

activities, the carer can have significant impacts on the person's ability to achieve his/her activities. The carer, for instance, might not be available at the times the mobility impaired person needs to attend university lectures. In this context, the actions of someone else are interfering with a person's agency, either expanding or reducing capabilities. To address this issue, Sen distinguishes two types of agency success (which means: one's goals become realised) (Keleher 2014): a) realised agency success and b) instrumental agency success. In the first case one's objectives are realised even if the individual doesn't play any role in the achievement. Following the above example, the choice of the carer to assist the mobility impaired person to visit the university beyond his/her normal working hours will result in a realised agency success for the mobility impaired person. In the second case, it is the actions of the person that facilitate the achievement. In our example, this concept could be understood if the mobility impaired person offered some sort of incentive for the carer as compensation for working beyond their normal working hours.

Sen intentionally kept the CA framework loose so that a variable can be considered as a functioning, capability or characteristic that influences the capabilities set, depending on the circumstance. However, this under-specification is a subject of criticism by researchers, who argue that it makes the application of the CA to practical, everyday problems difficult. These difficulties can be attributed to the ambiguities in defining capabilities as well as to its bottom-up approach, requiring participation of the people immediately involved (Comim 2008). Another issue of the CA that is commonly raised is that of individualism. As mentioned in section 2.5.1, the argument here is that the CA fails to account for capabilities that arise from collective participation since it is focused on the individual level. As such, there is the risk of omitting the importance of the institutions, public bodies and communities that helped create and sustain the capabilities in the first place (Deneulin 2008). However, some researchers (Alkire 2008) argue that such intrinsic importance to group capabilities should be viewed with caution since participation in collective activities on its own is not enough to evaluate well-being on the individual level. The Dial-A-Ride example mentioned in section 2.5.1 is an example of a group capability that doesn't necessarily result in increased well-being.

3.3 The capabilities approach in transportation literature

Given the recent interest of the CA in the context of accessibility, there are two strands of studies in the literature: Studies that focus on defining accessibility

through the use of the CA, and studies that use the CA framework to examine issues of equity in transport through specific case studies. It is important to note that these two bodies of research should not be viewed in isolation with each other as the concepts and theoretical foundations often overlap.

3.3.1 Capabilities approach in transportation literature: Definition of CA components

Hananel & Berechman (2016) argue that the first step towards translating the CA to the transportation domain is to define what is meant by capabilities. In their view, a combination of the extent of mobility and access to opportunities for individual population groups, especially the disadvantaged ones, could be considered as good candidates for capabilities. These capabilities should reflect the minimum conditions that allow the least advantaged groups to benefit from any transportation interventions. Thus, the functioning vectors may include measures such as the maximum allowable travel time, travel distance or travel expenses for all residents in the area of influence, focusing on the more disadvantaged. The authors conclude that the capabilities and functioning vectors should not be viewed as independent from one another, but should recognise and address the interactions between them. As an example of how this approach can be implemented in practise, Hananel & Berechman (2016) demonstrate that addressing the affordability of public transport through targeted interventions can result in the creation of capabilities for the less affluent people in King county, United States. However, although a step in the right direction, this example doesn't elaborate on the way such interventions can be expressed through the proposed functioning vectors, given the complex interactions of the defined capabilities.

In another study, Beyazit (2011) juxtaposed the core elements of the CA with concepts in transport research. In their analysis, functionings refer to the wider definition of accessibility as described in section 2.4. Particularly, the transport system constitutes the goods, while the provision of access to ones needs and wants is the functioning of the transport system. Travelling for leisure could be one of these functionings, as is travelling for social interaction. The capabilities then refer to the mobility element that enables people to move from one location to another physically, socially and financially, within a society and across societies. In this way, people possess a capabilities set which translates into an opportunities set of achievable functionings, from which they are free to chose. Manifestations of these choices could be the travelling mode or modes, the choice of locations, the reason to travel and the choice of travel time.

Hickman et al. (2017)'s interpretation of functionings and capabilities within

the transport context is similar to that of Beyazit (2011). In their view, the functionings represent what a person actually does and how. The realised functioning element is represented by the actual travel behaviour and participation in activities and as such, it is easier to measure. Measurement of capabilities on the other hand is more challenging. The authors propose an individual based accessibility definition that encompasses, alongside physical accessibility, issues such as the type of available infrastructure, land use, social and cultural norms and individual characteristics. The defined capabilities set is specific to each individual and reflects the freedom to choose from different potential functionings. However, this doesn't mean that two persons with similar functionings have the same capabilities. For example, a person with higher income may choose to have a similar mobility level to a person of a lower income by choosing not to own or use a car.

For Grengs (2015), functionings are translated as achievements of what a person manages to do or be. For example, using the different transportation options in order to be mobile, is an interpretation of a functioning. Having access to goods and resources can enable functionings, just as an unfavorable physical and social environment can disable them. Simply adding those functionings in a utilitarian way does not reflect their overall contribution to well-being since quality of life is also determined by the opportunities available for the individual. In this reading, a capability is a functioning an individual could have achieved. In this sense, accessibility as a measure of potential access to destinations reflects the notion of capabilities.

Pereira et al. (2017) proposes framing accessibility in terms of combined capabilities, having two separable but interacting components. This first one relates to a person's capability to access and use the transportation system, which depends on the interplay between personal and external factors. Personal factors may be individual characteristics such as physical and mental health, accumulated experience and financial resources. External factors may be the social environment as well as the transport system's design, price level information or availability. The second component refers to the more macroscopic view of accessibility which is related to the interaction between the transportation system and land-use patterns, and how this interaction acts as an enabler towards the expansion of capabilities. This includes elements of the transportation network such as network coverage and connectivity, as well as the spatial distribution of activities.

A somewhat different approach to defining capabilities and functionings is adopted by Nahmias-Biran et al. (2017) and Nahmias-Biran & Shiftan (2019) . In the author's view, the mobility element of accessibility represents the functionings in the transportation domain, while the ability to reach opportunities represent

capabilities. The former is associated with the act of travelling, the latter with the traditional definition of accessibility. The authors proposed the use of a logsum accessibility metric as a potential candidate for capturing the essence of capabilities, however since this indicator is calculated on actual travel patterns, it's usage contradicts the view of capability as the means to capture the possible opportunities that could have been chosen by an individual, reflecting the idea of 'freedom' in Sen's approach (Martens & Golub 2012)

Tyler (2006) approached accessibility through the CA following a more microscopic view. Capabilities are perceived as the combination of the individual abilities of a person, and the capabilities the environment provides. As Tyler (2006) states, the physical infrastructure might require someone to be able to step up 30cm to participate in an activity. If the person is not able to provide this capability based on their individual characteristics (eg. wheelchair user or elderly), then participation in the activity is not possible. Therefore, there is an interaction between what an individual can offer and what the environment can provide. The authors developed a measurement framework that relates the difficulty of achieving a task with the combination of required and provided capabilities (Holloway & Tyler 2013, Cepolina & Tyler 2004). The result determines whether a task is possible to achieve or not.

3.3.2 Capabilities approach in transportation literature: Case studies

Reviewing the most relevant literature was done by examining the scope of the studies with respect to equity issues in transport, the functioning vectors included, the definitions of capabilities and functionings used in the study, as well as the methodology followed and data used (table 3.1).

Authors	Scope	Factors included	Capabilities	Functionings	Methodology	Data
Hickman et al. (2017)	Investigating transport disadvantage between different income groups in Manilla, Philippines	Proximity to transport, security, air pollution, access to employment, income etc.	eg. Travel to work and other activities, Information, Natural environment	Similar to capabilities	Qualitative interviews with focus groups, self-disclosing desired and actual levels of PT experience	Online questionnaires, face-to-face interviews
Ryan et al. (2015)	Evaluating levels of interaction of PT for elderly people	Income, driving license, population density, gender, age, difficulties in boarding a bus etc.	The extent that elderly people can use public transport for the majority of their trips	Frequency of public transport use	Logistic regressions for capabilities and functionings using likert scale responses for dependent variable and factors as independent variables	Travel survey
Nordbakke (2013)	Investigating mobility of older women	Social networks, access to car, physical accessibility of the build environment, security	Availability of PT, availability of activities, access to information, access to alternative transport	Mobility levels	Qualitative interviews	Focus groups

Smith, Hirsch & Davis (2012)	Benchmarking transportation costs for rural communities	Income, accessibility to services, age and household composition etc.	Access and sustain of activities such as education, social participation, employment, healthcare etc.	Types and number of trips	Stratified sampling followed by qualitative interviews	Focus groups
Rashid et al. (2010)	Exploring transport disadvantage	Income, ethnicity, household composition etc.	Trip frequency, travel time, car dependency	Similar to capabilities	Principal component analysis followed by multicriteria evaluation	No information given
Maciel et al. (2015)	Exploring the mobility dimension of deprivation in Sao Paulo	Income, education, housing, access to information etc.	Mobility and accessibility	Commuting to work patterns	Generation of deprivation and accessibility indices	Census, travel diary data
Goodman et al. (2014)	Investigating the effects of providing free bus transport to young Londoners	Location, gender age, ethnicity, deprivation	Social participation	Free bus journeys	Qualitative interviews	Focus groups
Yang & Day (2016)	Effect of job relocation on travel wellbeing	Income, age, vehicle ownership, location, traffic etc.	Preferred PT mode	Used PT mode	SEM	Questionnaire

Chikaraishi et al. (2017)	Association between individual capabilities and travel time spent	Income, years of schooling, car ownership	Leisure (consumption) and employment (production)	Mobility	PCA derived index	Travel diary
Cao & Hickman (2019*a*)	Investigating desired and actual levels of participation to opportunities	Gender, education, age, employment, car ownership etc.	Travel safety, access to hospitals/groceries/education/work/recreation	Similar to capabilities	Multinomial regression	Travel surveys
Nahmias-Biran et al. (2017)	Evaluation of benefits of transport projects between population groups	Income	Access to different activities	Transportation modes used	Logsum accessibility indicator	Synthetic data from activity based models
Lira (2019)	Examining equity to accessing opportunities	Gender, age, disabilities, safety, access to information etc.	Extrapolation of functionings to basic capabilities	Transport mode, travel time, proximity to other transport users	Exploratory analysis using likert scale	Online questionnaire survey

Table 3.1: Reviewed literature.

Investigating transport disadvantage between different population groups in the city of Manilla, Philippines, Hickman et al. (2017) used qualitative structured interviews to assess the levels of capabilities and functionings experienced by different individuals. The study showed that the more affluent members of society experience higher levels of capabilities and functionings relative to lower income groups. This translates to transport infrastructure investments in the city affecting disproportionately people with lower income. In this context, the use of qualitative interviews provided direct input on the desired potential of individuals for various capabilities (eg. levels of stress while travelling, levels of accessibility to employment etc.), however, the nature of the research methods used makes it inapplicable for use within the context of passive mobility data.

In two similar studies, Nahmias-Biran et al. (2017) and Nahmias-Biran & Shiftan (2019) used data from activity based models to demonstrate the usefulness of the logsum accessibility measure within a CA framework. Using a synthetic binary logit choice model specification, the authors defined different thresholds of logsum values to represent sufficient levels of capabilities. Although the authors demonstrated the usefulness of the approach within the context of different policy evaluations between a simple "rich/poor" dichotomy of population groups, it is unclear how the developed methods can be transferred to real world applications given the complexity of individual activity/travel behaviour.

Looking at the mobility component of accessibility for elderly people, Ryan et al. (2015), approached capabilities as the outcome of an individual's mobility resources. In this way, the potential of an individual to use public transport constitutes an element of the capabilities set. Functionings are chosen by an individual from the elements of the capabilities set, which could be all the different transportation options. The definition of realised functionings as actual behaviour is in line with Pereira et al. (2017), Hickman et al. (2017) 's and Beyazit (2011)'s interpretation. The approach is demonstrated using a case study in Stockholm, Sweden. Two independent logistic regression models are applied to a travel demand questionnaire survey. The first one uses as a proxy for capabilities the potential to travel given mobility resources, while the second repeated the experiment with responses about the actual travel behaviour (functionings). The study highlighted the importance of attributes such as living with a partner, health, number of cars, education etc. in the ability to use public transport, however, the study focused only on the mobility element of accessibility, without providing insights on the levels of access (potential or actual) to particular opportunities at a destination.

A similar methodology was adopted by Cao & Hickman (2019*a*) in a study

aiming to understand the desired levels of participation in activities and the actual levels of participation in East Beijing, China. The authors used the former as an indication of capabilities and the latter as an indication of functionings. Using qualitative interviews, they assessed the difference between responses using standard ANOVA (Analysis Of Variance) and F-test statistical techniques. To determine whether spatial effects were significant in the analysis between different geographical districts of the case study area, the authors used a MNL regression model. The authors repeated the methodology in another case study for three east London district areas (Cao & Hickman 2019*b*). The authors were able to identify significant differences between capabilities and functionings across individuals of different socioeconomic backgrounds, however, evaluating complex mobility behaviour in this context is difficult using the proposed methodology (Zhu et al. 2018, Xie & Waller 2010, Yamamoto et al. 2004). This is particularly important in the context of passive mobility data characterised by individual trajectories.

Within a similar research scope, Nordbakke (2013) examined the mobility of elderly women in an urban setting. Capabilities are perceived as a combination of individual resources as well as contextual characteristics (eg. the wider socioeconomic environment). Barriers to mobility were perceived as disabling agents that constrict the capabilities set. The geographical scope of the study was the city of Oslo, Norway and the research method used was qualitative interviews and focus groups with questions designed to probe the potential of travel. The authors found that sociodemographic variables together with residential density were strong predictors for both the capability and functional elements of the study. Again here, the study focused on the mobility element of accessibility without demonstrating how the use of different transportation modes result in increased levels of access to opportunities.

Examining the equity gap for a case study in the city of Santiago, Chile, (Lira 2019) interpreted capabilities within the context of both available opportunities, the freedom to choose between them and the ability to convert resources into valuable functionings. All three components interact with each other in the context of performing the activities an individual might value. The authors highlighted the difference between this interpretation of evaluating equity and methods based on the needs, satisfaction, happiness or subjective well being commonly employed in the transportation equity literature. Using questionnaire surveys, the authors used the perceptions of individuals to construct 'weighted functionings' or functionings weighted according to the relative importance an individual is attributing to them. They then used this notion to 'probe' on the individual capability levels. In this sense, the study facilitated an exploratory approach to the notions of functionings

and weighted functionings, as opposed to providing a modelling methodology that can be transferred to different application settings.

In a study to identify minimum income standards for transport use within rural communities, (Smith, Hirsch & Davis 2012) used the CA to place income in the wider notion of well-being. Income however, is only one of the factors that affect people's capabilities to function. As a result, the authors extended the definition of minimum income to refer to all the goods, services, opportunities and choices to participate in society. This definition was then presented to focus groups that were free to describe the capabilities needed to attain this function. These were then translated to specific metrics such as types and numbers of trips required which, along with other datasets (accessibility indicators, cost per mile, distances etc.), were used to formulate the minimum income standard threshold for different rural population groups. Specific types of trips derived from this approach were: access to groceries, household services and goods, transport, education, health, employment, social and cultural participation. Similarly to other focus groups related studies reviewed in this section (see table 3.1), the research methods are out of scope in the context of using passive mobility data.

Other authors (Orr 2010) proposed to define the capabilities set by focusing on activities, both realised and potential. Once identifying these, the individual capabilities required to achieve these can be mapped out (eg. access to sufficient income). The authors then framed this approach in terms of evaluation of different transportation interventions aimed at minimising transport disadvantage and social exclusion for elderly and disabled people. They proposed to break down an activity into tasks and assess each task individually. For example, the activity "going to a shop" has a set of necessary tasks embedded, one which could be "taking the bus". Specific barriers can then be associated with the particular tasks, such as "fear of crime walking to the bus stop". In contrast to the above mentioned case studies, this approach to defining and measuring the capabilities set is inherently data driven. The authors suggested possible sources of mobility data for these tasks such as GPS. However, that implies having a robust methodology to infer specific tasks/activities from unlabelled mobility data. As such, case studies where this approach is applied were not provided.

A quantitative approach to evaluating transport disadvantage using the CA was adopted by Rashid et al. (2010). In this study, the authors used the CA to identify a set of variables believed to influence functionings, such as low trip frequency, long distance travel, travel time and high private vehicle use. These variables were related to three factors: socio-economic characteristics (e.g. income, ethnicity), land-use characteristics (e.g. population density, neighbourhood

types) and public transport characteristics (e.g. routes, stops). The next step of the process was to apply dimensionality reduction techniques to define a set of indicators from the input variables and associate them with the functioning variables by a linear additive relationship. Finally, the authors proposed performing multi-criteria evaluation techniques to different transport disadvantage scenarios and examine the distribution of the indicators. Although the methodological framework provided by the authors draws elements from the CA to identify and include factors related to transport disadvantage, subsequent analysis doesn't differentiate between capabilities and functionings.

3.3.3 Capabilities approach in transportation literature: Discussion

The CA has been applied to a wide range of social issues in transport, ranging from investigating the impact of specific transport interventions to evaluating transport related social exclusion. In nearly all cases, the studies were based on empirical findings within a specific geographical context, while the focus was on disadvantaged groups (eg. low income people, elderly, slum dwellers etc.) and within a comparative evaluation framework. A considerable proportion of the reviewed studies were qualitative, in line with the body of literature covering social aspects of transport (Lucas & Porter 2016). The ones that were more quantitatively oriented used statistical tools such as Structural Equation Models (SEM), Principal Component Analysis (PCA) , logistic and MNL regression models. This suggests that there is currently no consensus among researchers on how to quantitatively operationalise the CA for issues related to transport and social exclusion. This trait is not an exclusive property of the CA, and is associated with the nature of considering equity in accessibility evaluations as discussed in section 2.4.2.

The definition of the elements included in the capabilities set and the corresponding functionings is used interchangeably in some studies. This is not uncommon and has been identified in applications of the CA to other social aspects beyond transport (such as quality of life) (Robeyns 2005*b*). Reasons for this can be traced in the definition of functionings as enablers to achieve the defined capabilities, but also the close relationship between transport concepts such as mobility and accessibility (for example, mobility can be considered both a functioning (using the bus) and a capability (ability to move)) (Chikaraishi 2017). In all cases, however, there is a distinction between what is measured (functionings) and the hypotheses to be tested (capabilities).

On the other hand, there exists a general consensus on the factors influencing the capabilities set. This includes either focusing on the socioeconomic character-

istics of an individual, the wider environment (both physical and social) or both. In line with the social exclusion definition as provided by the Social Exclusion Unit (Social Exclusion Unit 2003), sociodemographic variables such as income, age and gender are all defining factors that influence accessibility and have been included in the majority of the studies. Variables of the wider social environment such as deprivation, although not explicitly accounted for, have been taken into account during the design phase of most of the reviewed studies. Physical characteristics such as distance to amenities and density of public transport have also been adopted by the majority of the studies as important factors that shape the capabilities set.

In terms of data used, qualitative data acquired using the CA as a methodological basis has been the most widely used. Such data provide a deeper insight into the reasons behind accessibility issues as expressed in the latent capabilities set, particularly in the cases where actual choice behaviour is restricted in some way. However, there is currently no formal instrument designed to obtain information on an individual's choices while avoiding the risk of subjectivity and bias in the responses, particularly when incentives for participants to provide accurate responses are insufficient (Krishnakumar 2013). Furthermore, due to the complex and costly nature of data collection, scaling the findings to the population level is difficult. This is of particular importance when evaluations of differences across population groups is the goal of the study. In terms of quantitative studies, repurposed data (such as travel surveys) have been used by some of the studies. Such data don't provide any insights on the capability sets, and they mostly relate to data on achieved functionings. As such, it is up to the researcher to define the latent (hidden) capabilities set using previous research in the field or additional research through focus groups. The main advantage is the larger penetration of the general population, which is often coupled with sociodemographic data at an individual or household level.

Although all of the studies reviewed here make implicit and explicit connections of the CA with accessibility as the multidimensional concept defined in section 2.4, none elaborate on how existing numerical accessibility methods can be used within the CA framework.

Finally, in spite of the advantages of passively generated mobility data from transport service providers, namely larger samples, regular update rate, low cost and the potential for longitudinal studies (Pelletier et al. 2011, Bagchi & White 2005), none of the reviewed literature has explored their potential to extract quantifiable evidence of social exclusion and transport disadvantage. This is true within the accessibility literature in general (Anda et al. 2017) and the CA in particu-

lar. This is largely due to the unlabelled nature of such datasets, requiring an additional step to infer activity types at a destination.

3.4 How do existing numerical accessibility measures fit within the Capabilities Approach framework?

After examining the way the CA has been used in transport studies, some common themes emerge between accessibility and the CA:

- Capabilities represent the potential of an individual to reach and engage with opportunities. Realised functionings represent the observed behaviour of the above. Both of the terms are in line with the general definition of accessibility as set out in section 2.4.
- The focus of the CA is the individual and in this sense in line with individual based accessibility indicators. Moreover, it takes into consideration the influence of internal and external factors that shape the individual capabilities set.
- The capabilities set is not static but in constant interaction with the components that shape it and the realised behaviour expressed by the actual functionings. The evolving nature of the capabilities set extends both spatially and temporally, in the sense that is modified based on location and time.

Moreover, there is a causal structure between the factors that shape the capabilities, the capabilities themselves and the functionings. This causal structure appears to be hierarchical in nature, with the functionings appearing at the bottom of the hierarchy and the factors appearing at the top.

The fact that the CA framework is applied at the individual level renders accessibility as the potential of interaction and cumulative based indicators incompatible (Pereira et al. 2017). This is because such measures analyse accessibility at a particular location where it is assumed that all individuals at the location's catchment have the same accessibility levels (Geurs & Ritsema van Eck 2001). This is true even if disaggregation for different socio demographic groups occurs. However, it is important to notice that the concept of potential access to opportunities represented in such indicators (particularly cumulative based accessibility indicators) to express freedom of choice, is still relevant in the context of defining capabilities.

On the other hand, individual accessibility indicators approach the concept as an attribute of an individual. This property is in line with the general premises of the CA framework. This property is in line with the general premises of the CA framework. However, as both utility and space-time accessibility use actual individual behaviour to derive the indicators, it is important to emphasize that the concept of freedom in Sen's approach is not adequately represented.

A discussion on the strengths and weaknesses of each accessibility indicator was given in section 2.2.1. The following section will relate these insights in the context of using the CA as a platform to express equity issues in transport, focusing on individual accessibility indicators.

3.4.1 Space-time accessibility indicators

Space time accessibility focuses on the potential area of opportunities that can be reached within a given temporal constraint (Geurs & Ritsema van Eck 2001). These potential areas represent the accessibility levels of an individual and are derived from individual level mobility data (such as GPS traces, travel diaries etc.). They are calculated using either the volume of the space-time prism or its projected area on the spatial plane.

In its simplest form, the shape of the construct is purely geometrically derived based on the set of locations that can be reached within a predefined time interval, assuming some speed of movement (Ettema & Timmermans 2007). However, many authors have extended the framework to account for different assumptions. For example, it is not uncommon to account for the relative attractiveness of the different opportunities based on some utility or gravity type based formulation (Geurs & Ritsema van Eck 2001). Other authors have examined how different assumptions about the uncertainty of the scheduled times between activities can be incorporated in the standard space-time prism calculation (Ettema & Timmermans 2007). Focusing on quantifying the uncertainty of the traveller's movement in space, Winter & Yin (2011) formulated a probabilistic approach for calculating the potential path's surface. The uncertainty in this case is evaluated using stochastic processes, such as random walks, constructed using unbiased or biased transition matrices. The latter case is useful for accounting for the relative attractiveness of destinations.

In all cases, however, the space-time prism is determined by the spatiotemporal constraints such as maximum speed, distance, time budget or, in the case of introducing a bias, the attractiveness of a destination. Individual characteristics and external factor influences on accessibility can only be determined by looking at differences between space-time derived metrics across individuals of different

sociodemographic characteristics. For example, Kwan (1999) compared the potential path areas between individuals of different gender in Ohio, USA, when assessing the relative accessibility to different day-to-day opportunities (such as shopping, entertainment, education etc.). In a study focusing on children's access to urban opportunities (such as home, school, recreational) Casas et al. (2009) postulated that the resulting space-time metrics could be used to reflect household characteristics such as income and the guardian's working hours.

Following from the above discussion, space-time accessibility measures, although useful for different applications ranging from assessing policy implications to visualisation (Neutens et al. 2011), do not quantify the magnitude and significance of interactions between the variables. Moreover, the causal structure between resources/capabilities and functionings is difficult to represent using space-time accessibility measures.

3.4.2 Utility based accessibility indicators

Utility based accessibility measures assign a utility to each destination choice element from a finite destination set. Within a multinomial logit/logsum formulation, the destination set can be represented as a multinomial distribution over a finite set of opportunity categories such that for each $x_i \in \{0...n\}$ destinations belonging to $i \in \{1...\kappa\}$ opportunity categories, there is a corresponding probability vector $p_1...p_\kappa$ with $\sum_\kappa p = 1$. This probability vector is informed by the utility assigned to each x_i and the probability vector is inferred from the individual's observed behaviour. Although the resulting probability vector can relay to concepts of the CA as a representation of the capabilities set, such an assertion is not completely accurate as its calculation is not based in the utility travellers are experiencing, but on the utility of their final decision (Chorus & De Jong 2011). More complicated models including latent components for choice formation, could potentially be used to add a layer of abstraction between the observed choices and the hidden combinations of choice sets to represent the capabilities set (e.g. Cross-nested Logit (Vovsha 1997), GenL (Swait 2001)). These models generally perceive the latent choice sets as subsets of a 'master' set each one having a certain probability of being the 'true' choice set. However, as already mentioned in section 2.2.1.4 these structures were originally introduced to capture the correlations between different choices, resulting from incomplete knowledge about the decision making process or captivity effects (choices not available to an individual by default), and not as an alternative to the reasoning process that defines the potential behaviour of an individual. On the other hand, incorporating a degree of prior belief in a Bayesian setting to inform an individual's decision making process has been

proposed as a more flexible alternative to account for the behaviour uncertainty (Brownstone 2001, Daziano et al. 2013). This stems from the fact that using such a specification, uncertainty and additional decision making process assumptions can be accounted for within the context of each potential choice. This adds more flexibility for representing the shape of the capabilities set as in this specification, data that are not directly related to the utility function can be included. Examples include prior beliefs made on the basis of past studies or subject matter expertise that can express the potential of an individual's behaviour before observing the data.

In its basic form, the utility function can be formulated to include explanatory variables in the form of covariates that are assumed to contribute to the choice of κ alternatives. This is usually approached in a deterministic way, as a linear additive function of covariates with coefficients β_κ: $U_i = V_i + \epsilon_i = \sum_\kappa \beta_\kappa x_{i\kappa} + \epsilon$ where ϵ corresponds to the error term. The covariates $x_{i\kappa}$ can be chosen to represent the attractiveness of different destinations as well as personal characteristics of travelers (Dong et al. 2006). In this way, external and internal factors that influence the accessibility of an individual can be included in the model. For example, Bocarejo S & Oviedo H (2012) formulated the utility function using different variables such as occupation, age and income level as well as land use variables such as activities distribution. The goal of the study was to analyse the effects of the introduction of a new bus rapid transit line on elements of social exclusion for the travellers affected. In another study Doi et al. (2008) used a utility-based accessibility model to investigate the travel patterns of elderly people and their relationship with gender and ethnicity in Japan. The linear additive form in the deterministic part of utility function is considered a standard in discrete choice modelling literature, and this holds true for simple to more complicated models. Such a specification introduces restrictions in the context of the CA as representation of complex interactions between the elements of the functioning vectors becomes difficult. Although there exist non-linear representations of utility parameters, estimation, model convergence and interpretation is more difficult (Train 2009).

In the context of CA, the functioning vectors have been identified as the set of resources, the socio-spatial context, and preference mechanisms that shape the capabilities set. These factors are often context and situation dependent and influence the capacity of an individual to choose an achievable functioning from the capabilities set at different levels (e.g. long term, short term). Utility based models are in essence discriminative models capturing the choice probability of the utility function given the different explanatory variables $P(U|x)$, which in

the context of CA can be related to the functioning vectors. Inference using observed data results in estimates of the strength and significance of the covariates on the choice set. Although this process can provide insights on the degree of influence of the explanatory variables in the probability choice vector, it generally doesn't impose any specific structure on the relationship between the explanatory variables and the different opportunity choices apart from the assumption that choice probability is expressed by a linear combination of covariates. Approaching a latent quantity such as capabilities using an approach that does not impose a generative structure makes interpretation of the interactions difficult. On the other hand, using a causal structure on the functioning vectors would allow easier interpretation of the model's output. Causal modelling techniques such as belief networks, Bayesian and decision networks can be a promising alternative in this regard. Such modelling specifications have been applied successfully in problems where modelling decision making process under uncertainty is the focus. Examples include transportation mode choice behaviour (Ma 2015, Verhoeven et al. 2005, Daziano et al. 2013), location choice modelling (Ma & Klein 2018) and activity-travel behaviour (Koushik et al. 2020). The CA in this regard can provide the basis upon which the causal structure is defined.

3.5 A Capabilities Approach accessibility framework

Following from the discussion of sections 3.3 and 3.4, it is clear that no individual accessibility numerical measurement framework exists that captures all the elements of the CA as set out in section 3.3.1. Space-time accessibility measures can be related to the concept of capabilities through the use of the space-time prism. However, its computation emphasizes primarily on the spatiotemporal components of accessibility. On the other hand, utility-based measures focus on measuring the actual behaviour (the functionings) and are not suited to quantify the potential behaviour of individuals. When it comes to including all the factors that influence the shape of the capabilities set, utility-based measures can incorporate external and internal factors through the utility function however the linear functional form limits the flexibility of expressing complex interactions in an interpretable way. Space-time measures on the other hand, do not provide a native framework to do so. However, they can be modified to include destination attraction factors through incorporating a utility function in the calculation of the space time prism (Geurs & Ritsema van Eck 2001). In both individual accessibility measures, interactions between all the factors influencing accessibility are not explicitly quantified. As a result, researchers interested in the interactions

between accessibility and personal characteristics often use other modelling frameworks such as structural equation modelling (Simma & Axhausen 2003, Van Acker et al. 2010, 2007).

3.5.1 The Capabilities Approach to accessibility as a graph

Using the CA framework as a guide, it is possible to form a basic structure that specifies how the basic accessibility components relate to each other and interact in a causal way. The different concepts of the CA and the way they are linked are shown in figure 3.1 (Mitra 2006, Beyazit 2011). At an observable level, one encounters the functionings of an individual. Within an accessibility setting, this node refers to the realised activities as well as the realised transportation modes used to reach those activities. Moving one level up the hierarchy there exists the latent set of capabilities that form the choice set of an individual. These are all the potential opportunities an individual could choose. In this setting, a realisation of a chosen element of the capabilities set leads to an observed functioning. This in turn is influenced by personal, environmental and social characteristics, as well as the commodities a person has in his/her possession. All variables of this representation are expressed through stochastic quantities that aim to quantify uncertainty from incomplete knowledge about the state of variables, as in the case of capabilities, or from noisy and erroneous measurements, as in the case of functionings. In this line of argument, the CA is used as a theoretical framework upon which an interpretation of individual accessibility is constructed, as opposed to providing the platform upon which issues of social equity are discussed. This is in accordance with the current practice of the CA used in empirical and applied studies (Robeyns 2005*a*). Nevertheless, by using the hierarchical structure between personal/environmental characteristics, capabilities and functionings, the degree of contribution of each of the components to an individual's levels of social exclusion and transport disadvantage can be evaluated.

The process is relevant for each individual and takes place in space and time during the act of reaching opportunities. In this setting, the capabilities set changes depending on the characteristics of the environment that exists in each location at a particular point in time ($t = 1...n$). This representation imposes a structure on accessibility through the use of a directed graph, where the nodes represent the components of the CA and the edges the relationship between them. The graph is acyclic, in the sense that no closed loops appear between the nodes. This allows information to flow from the top level to the bottom level nodes. The whole process should not be independent between subsequent time steps but should capture the dynamic evolution of capabilities through time.

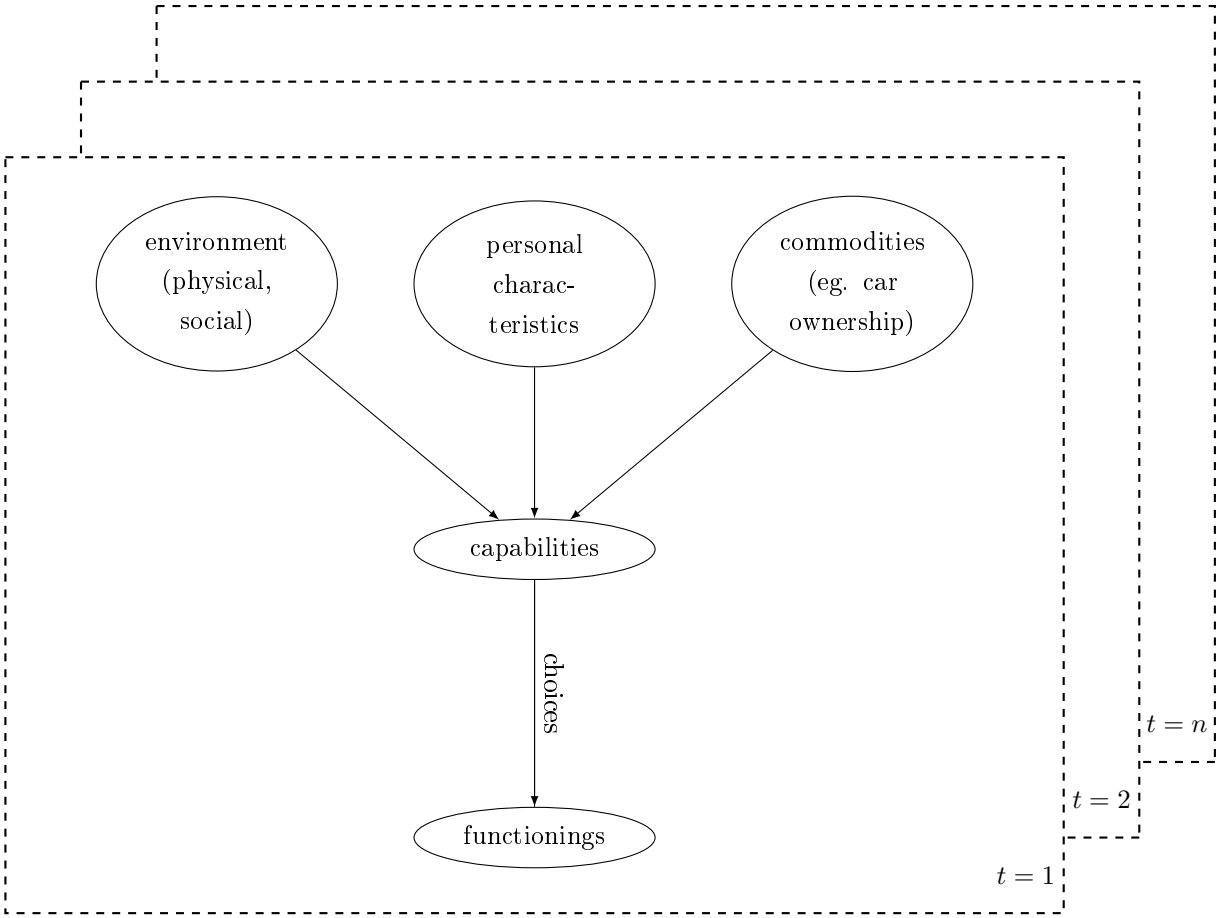


Figure 3.1: The CA (adopted from Mitra (2006)).

3.6 Chapter summary

This chapter provided an overview of the CA, with emphasis placed on the relationship with accessibility. It is argued that, given the themes most commonly mentioned in the literature on this relationship, none of the existing numerical accessibility measures succeeds in providing a framework for implementation. Drawing from the hierarchical and causal structure of this formulation, it is postulated that graphical models could provide the mathematical foundations to represent the building blocks for the Capabilities Approach to Accessibility (CAA) framework.

The next chapter will introduce the background on the theory behind graphical models as well as their applications within the wider accessibility/transportation literature.

Chapter 4

Graphical Models

4.1 Chapter overview

The previous chapters discussed some theoretic aspects of the link between the Capabilities Approach (CA) and accessibility. It was argued that traditional approaches such as discrete choice models are insufficient given the requirements of this thesis (section 3.4). Therefore it was postulated that graphical models could provide an implementation alternative to already existing accessibility measures.

This chapter will introduce some basic terminology of graphical models before proceeding to the description of two main families of models used within the wider accessibility/transportation literature: probabilistic graphical models and structural equation models. Given both the inferential requirements from unlabelled mobility data and the accessibility modelling requirements of this thesis, each main type of model will be assessed through the lens of computational intelligence (inferring semantic information from unlabelled mobility data) and statistical inference (inferring the properties of the underlying model and interpreting the relationship between variables).

4.2 Graphs

Graphical models are statistical/mathematical models that use a graph structure to represent the relationship of a set of variables. Such models have been used throughout an extremely wide domain of disciplines, with applications ranging from psychology, social sciences and econometrics to genetics and machine learning (Koller & Friedman 2009, Pearl 2009).

More general, in discrete mathematics, a graph G is an object consisting of nodes and edges such that $G = (N, E)$ where N is the set of nodes $N = \{\mathcal{X}_1, \mathcal{X}_2 ... \mathcal{X}_n\}$ where $\mathcal{X}_i$ can correspond to a mathematical object (eg. random

Table 4.1: Basic terminology of graphs

Term	**Description**
Nodes	The set of random variables $N = \{\mathcal{X}_1, \mathcal{X}_2 ... \mathcal{X}_n\}$
Edges	The set of pairs where each pair is either $\mathcal{X}_i \rightarrow \mathcal{X}_j$, $\mathcal{X}_j \rightarrow \mathcal{X}_i$ or $\mathcal{X}_i - \mathcal{X}_j$
Subgraph	A subset G_i such that $G_i \subset G$
Directed edge	An edge connecting nodes that have a parent/child relationship $\mathcal{X}_i \rightarrow \mathcal{X}_j$
Undirected edge	An edge connecting nodes that are neighbors such that $\mathcal{X}_i - \mathcal{X}_j$
Path	A sequence of edges that connect a sequence of nodes. These can be directed or undirected. The term trail instead of path is often used for undirected paths.
Neighborhood (of a node)	The immediate adjacent nodes of a particular node that are connected with an edge.
Cycle	A cyclic path occurs in a directed graph when a node leads to itself: eg. for $N = \{\mathcal{X}_1, \mathcal{X}_2 ... \mathcal{X}_\kappa\}, \mathcal{X}_1 = \mathcal{X}_\kappa$.
Directed Acyclic Graph (DAG)	A graph is acyclic when there are no cycles occurring. Such a graph is referred to as a directed acyclic graph .

variables, deterministic functions, linear models) and E is the set of edges connecting the nodes.

There are two types of edges, directed and undirected. A directed edge between $\mathcal{X}_i \rightarrow \mathcal{X}_j$ implies a conditional probability $P(\mathcal{X}_j|\mathcal{X}_i)$ while an undirected edge $\mathcal{X}_i - \mathcal{X}_j$ implies a joint probability $P(\mathcal{X}_i, \mathcal{X}_j)$. Table 4.1 summarises the basic terminology of graphs in the context of graphical models after Koller & Friedman (2009).

As the type of graph is determined by the type of edges, one can make the distinction between directed and undirected graphical models. However, there exists a structure where both directed and undirected edges exist in the same graph, the mixed graphical models.

4.3 Graphical models

In the context of using the Capabilities Approach (CA) as a structure to model accessibility for applications related to transport related social exclusion and equity, graphical models offer some distinct advantages compared to traditional accessibility indicators:

- The graphical structure encodes causal relationships between a set of stochastic variables for a problem domain. In this way, the causal flow between functioning vectors/capabilities /realised functionings can be explicitly represented.
- The conditional probabilities used to express relationships, allow for a fully probabilistic approach to the problem. In the context of representing the potential behaviour of individuals, this enables incorporation of assumptions (derived by secondary data and prior knowledge) related to forming the latent capabilities component as a set of conditional probabilities
- Expressing every variable in the model as a stochastic variable, enables quantification of uncertainty through the shape of the resulting probability distributions. This is highly desirable in the context of incomplete information about the factors influencing capabilities.
- Graphical models allow propagation of information among the variables, as dictated by the graphical structure. This allows updating the state of capabilities related variables in the face of new evidence, enabling a more precise representation of capabilities as a compromise between resources (expressed by prior information and covariates) and welfare (expressed by the realised functionings).
- Graphical models can be easily extended to account for the dynamic representation of processes in both the temporal and spatial domain, rendering them ideal for modelling evolving phenomena such as the capabilities of an individual.

4.4 Probabilistic graphical models

Probabilistic graphical models first appeared in the area of statistical physics in 1902 when modelling the interactions between particles (Gibbs 2014). Two decades later, Wright (1921) used such models to represent the genetic inheritance

of species. More recently, probabilistic graphical models were the key to understanding the human genome through DNA sequencing (National Human Genome Research Institute 2017) and today they are an integral part of many machine learning algorithms.

In probabilistic graphical models, the model is represented by a joint probability density over the set of random variables, while the graph structure represents the skeleton dictating how the variables are related with each other. This graphical representation has several benefits (Koller & Friedman 2009):

- It allows the representation of a phenomenon using a model structure that captures the flow of information in a causal way
- It significantly reduces the inference feature space by a set of probabilistic assumptions on the state of dependencies between the variables, making inference tractable.
- It provides a transparent framework for humans to understand and evaluate the semantics and properties of the phenomenon associated with the model.
- It allows for knowledge discovery, through the use of queries and what if questions through the use of conditional probabilities.

There are two main probabilistic graphical model types: undirected graphical models and Bayesian networks.

4.4.1 Undirected Graphical Models

As the name suggests, an Undirected Graphical Model (UGM) is a representation of a joint distribution between random variables where the factorization relative to the graph structure doesn't assume any direction. Figure 4.1 shows a simple model with four random variables.

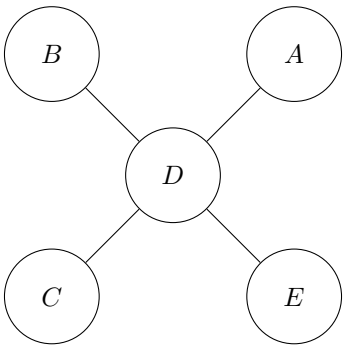


Figure 4.1: A simple Undirected Graphical Model

Such models are often referred to as Markov networks or Markov Random Fields (MRFs) due to the 'memoryless' property of Markov processes. This dictates that information about the events that are separated by a random variable not belonging in their immediate neighborhood, are conditionally independent given the separating node. In the graph of figure 4.1 this translates to $P(C \perp A|D)$. This does not imply complete independence between C and A, it simply transcribes the fact that if we know the outcome of D then the outcome of A does not contribute to the outcome of D as all information related to C is determined by D. On the other hand, if the outcome of D is not known, then information about A could help determining C.

The joint distribution of an undirected graphical model is encoded by a set of factor potentials which are determined based on the graph structure. In the graph of figure 4.1 the factor potentials are $\phi_1(C, D), \phi_2(A, D), \phi_3(B, D), \phi_4(D, E)$. The term potential intuitively refers to the compatibility among variables in their immediate neighborhood: the larger the potential the more compatible are the variables in the graph configuration (Liao, Fox & Kautz 2007). The joint density in the above example is then:

$$P(A, B, C, D) = \frac{1}{Z}\phi_1(C, D), \phi_2(A, D), \phi_3(B, D), \phi_4(D, E) \tag{4.1}$$

where, in the case of discrete random variables,

$$Z = \sum_{A,B,C,D,E} \phi_1(C, D), \phi_2(A, D), \phi_3(B, D), \phi_4(D, E) \tag{4.2}$$

is the normalising constant transforming the equation 4.1 into a valid probability distribution, and is referred to as the partition function.

Advantages of such specification is that it allows greater flexibility when representing the interactions between the variables, as no assumptions on the direction of flow of the information is made. However, this can have an impact of the interpretability of the results as it is not clear what the cause and effect relationship is for the analyst (Koller & Friedman 2009).

Of the multitude of undirected graphical model specifications (Robert 2014), only the ones that were used in applications related to the broader accessibility/transportation literature are discussed.

4.4.1.1 MRFs in statistical inference

Ubiquitous examples of this class of graphical models for different disciplines are Gaussian Markov Random Fields (GMRF). A GMRF is a finite dimensional ran-

dom vector following a multivariate Gaussian distribution (Rue & Held 2005). GMRFs are specified through a set of conditional distributions of one node given its neighborhood (Gelfand et al. 2010) using an undirected graph structure. For a random vector $\mathbf{X} = (X_1, X_2...X_n)$ with respect to a graph $G = (V_{1...n}, E)$ a GMRF has a probability density of the form:

$$P(X) = (2\pi)^{-n/2}|\mathbf{Q}|^{1/2}exp(-\frac{1}{2}((\mathbf{x}-\mu)^T\mathbf{Q}(\mathbf{x}-\mu)) \tag{4.3}$$

where $\mathbf{Q}$ is a positive definite precision matrix (the inverse of the covariance matrix) with individual entries $Q_{ij} \neq 0$ for $i, j \in E$ and $i \neq j$.

GMRFs are used in many fields of discrete spatial statistics when a task requires reasoning under uncertainty for different spatial phenomena, ranging from spatial econometrics to disease mapping. A common way to incorporate this specification to analytical and predictive models is through a regression framework. This can be done by using the spatial structure specified by the graph to account for the interdependence of observations due to their locations in space. In the static case, where no temporal dependency is assumed, Simultaneous Autoregressive Model (SAR) (Anselin 2013) and Conditional Autoregressive Models (CAR) (Besag et al. 1991) have been extensively used in different applications. Assuming an observation random vector $\mathbf{Y} = (Y_1, Y_2, ...Y_n)$ with variables representing the nodes of an undirected graph and $\mathbf{W}$ a matrix encoding the adjacency structure as specified by the edge connectivity between the variables in the graph then in the SAR approach:

$$\begin{aligned} Y &\sim N(0, C) \\ C &= [(I_n - \rho W)^T(I_n - \rho W)]^{-1} \end{aligned} \tag{4.4}$$

In this specification, ρ is a spatial coefficient controlling the strength of the spatial dependency while W causes simultaneous autoregression of each random variable on its neighbors (Hoef et al. 2017). On the other hand, within a CAR model specification, every variable of the random field is conditionally specified only by its neighboring nodes. In its simplest form (Intrinsic Conditional Autoregressive Model, ICAR):

$$Y_i|y_{-i} \sim N(\sum_{c_{ij}\neq 0} c_{ij}y_j, m_{ij}) \tag{4.5}$$

where y_{-i} represents all the realization's of Y neighbours, c_{ij} are the elements of the spatial dependency matrix and m_{ij} is the i, j element of a variance matrix $\mathbf{M}$.

Both models have been used in applications within the broader transportation literature, particularly with applications related to modelling the interactions between land use and transportation. The choice of one over the other depends on the type of assumptions on the degree of influence of the spatial configuration, as well as the particular goal of the study. When observations are assumed to be correlated on global level, a SAR specification is more appropriate. When the effect of the spatial configuration is assumed to be localised, a CAR model can be used instead.

For example, following the assumption that the spatial configuration of residential properties locations is correlated with their price, Löchl & Axhausen (2010) used a SAR specification to model the influence of land use and transport characteristics in property prices. In another study, Ahlfeldt (2011) used a similar model to explain the influence of land use in property prices within a gravity-based accessibility model. They found that accounting for the spatial heterogeneity of land value can significantly improve standard gravity based employment/transport accessibility models. Investigating the hypothesis that jobs that can be accessed by different transportation modes between neighboring districts interact spatially with each other (due to similar physical and socioeconomic conditions), Wang & Chen (2015) used a SAR model to test the significance of the spatial effect of gravity type accessibility measures for walking, public transport and bus regressed on socioeconomic and built environment characteristics. Implementation of such a model informed the discovery of the transport disadvantage of single-parent households on transport-based job access. Looking at the travelling behaviour of disabled people, Bantis et al. (2017) used a CAR model to identify the most likely transportation access points that people might use in case of an emergency. They found that the spatial variability accounted by a Poisson regression with a CAR specification can account for the increased uncertainty of people's whereabouts during such events. Modelling the subjective travel satisfaction of travelers using sociodemographic attributes, location variables and travel related characteristics, Dong et al. (2016) used a CAR specification to account for the spatial effects within a multilevel modelling regression framework for the city of Beijing, China. They found that modelling the spatial correlation between district level covariates, can help prioritise potential transportation interventions between districts.

4.4.1.2 MRFs in computational intelligence

Besides tasks related to statistical inference, MRF have also been used for computational intelligence in machine learning algorithms. This broader category of MRFs is generally concerned with the efficient representation of the underlying

properties of data and also in the task of generating accurate and reliable predictions (Karlaftis & Vlahogianni 2011). Researchers using such models are mostly interested in data driven classification/clustering tasks given a highly non-linear feature space. Due to this shift of focus, applications directly related in accessibility topics that use MRF in such a way is more limited compared to tasks inference tasks. On the other hand, applications of MRF through computational approach allow the exploration of mobility related data that can reveal a great deal of detail related to individual's mobility behaviour. For example, Liao et al. (2006) used a Relational Markov Network (RMN) to classify raw GPS traces to specific activities such as driving, working and using public transportation. A RMM is a MRF that extents factor potentials into higher hierarchical layers describing how they are linked with each other:

$$P(Y|X) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \prod_{v_C \in C} \psi(C)\phi(v_C) \tag{4.6}$$

where $\mathcal{C}$ is the set of all higher level factor potential cliques and v_C is the set of all factor potentials defined by the graph.

In this way, the authors were able to use the spatiotemporal structure of individual GPS traces together with secondary information to discover the actual activity profiles of individuals. In a later study, the authors extended the hierarchical structure of this model to to include clustering at a factor potential clique level to discover contextual significant activity places such as visiting a friend, working etc.(Liao, Fox & Kautz 2007). Using a spatially coarser cell phone CDR (Call Detail Records) dataset, Widhalm et al. (2015) used a RMN to cluster the raw data into activity classes such as home, work, shopping and leisure. They did that by defining a set of clique potentials representing joint densities of of combinations of variables such as activity type and land use, activity type, starting time and duration of activity etc. They found that such an approach is able to efficiently reconstruct activity patterns appearing in conventional travel surveys (such as paper and pencil surveys, computer assisted telephone surveys etc.).

Modelling the higher level of interactions between different activities, Markov Logic Networks (MLNs) have been proposed as a promising technique (Yang 2009). A MLN is a representation of first-logic arguments and their relationships within an undirected graph structure (Murphy 2012). The arguments are referred to as formulas and are constructed using four types of symbols: constants, variables, functions, and predicates. Constant symbols represent the constant objects in an assertion, variable symbols allow to range over the objects, function symbols map objects to other objects in the domain and predicate symbols represent the rela-

tions of objects in a domain. Examples of such arguments could be (Li et al. 2017): The rainy weather causes car choice preference to increase, the rainy weather causes traffic jams, traffic jam causes public transit to have a longer travel time. In terms of applications, Li et al. (2017) followed a MLNs approach to modelling transportation choice behaviour. The logic arguments used in the study related transportation mode choice with levels of satisfaction using data from stated preferences survey. They found that such a framework can offer a richer representation of unobserved uncertainty in transportation mode choice modelling compared to a multinomial logit model.

Discovery of the choice of transportation modes used was also the research topic of Zheng et al. (2008). The authors used GPS traces were to infer whether an individual is walking, driving, using the bus or cycling. Using Conditional Random Fields (CRFs), the authors were able to capture the spatiotemporal dependence between the transportation modes and the raw observations. A CRF is a specific type of MRF where all factor potentials are conditioned on input features (Murphy 2012):

$$P(Y|X,w) = \frac{1}{Z}\prod_{c}\psi(y_c|x,w) \tag{4.7}$$

where $\psi(y_c|x,w)$ is usually modelled as $\psi(y_c|x,w) = exp(w_c^T\phi(x,y_c))$, x are the observed variables, y_c the transportation mode labels and w the associated weight vector.

The authors commented on the potential of CRFs to model data of sequential nature, however, they found that the relatively unchanged transportation mode state spaces (as most of the individuals didn't change transportation mode within a trip) didn't allow fully leveraging the advantages of CRFs on labelling raw mobility data.

In another study, Mohamed et al. (2014) used Hidden Markov Random Field (HMRF) to cluster people travel patterns using AFC data and socioeconomic variables. A HMRF is a generalisation of a Hidden Markov Model (HMM) which is a stochastic processes defined by a markov chain with latent (non-observable) states. The observable underlying graph within a HMRF specification is defined by a MRF:

$$p(y_i|x_i, x_{\mathcal{N}}) = p(y_i|x_i)p(x_i|x_{\mathcal{N}_\rangle}) \tag{4.8}$$

where $x_{\mathcal{N}_\rangle}$ is the neighborhood configuration of x_i. Using this specification, the authors were able to discover the varying travel to employment patterns and attribute them to the similarity of socioeconomic characteristics of different population groups.

4.4.2 Bayesian Networks

Contrary to UGMs, Bayesian Networks (BNs) factorise the joint probability density of a set of random variables according to the directional structure of a directed graph. Figure 4.2 shows a directed equivalent model of figure 4.1:

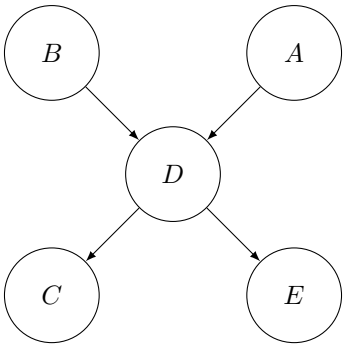


Figure 4.2: A simple Bayesian network

A BN is a DAG representing the conditional independence assumptions that factorise the joint distribution by the type and nature of connections between the variables. The joint density of figure 4.2 factors:

$$P(A, B, C, D) = P(C|D)P(E|D)P(D|B, A)P(B)P(A) \tag{4.9}$$

Bayesian networks combine the graph structure of a directed graph with the advantages of Bayesian inference. The term "Bayesian networks" doesn't necessarily imply that BN are committed to Bayesian statistics (Murphy 2012). Rather, they use the notion of conditional probabilities as specified by Bayes rule:

$$P(A|B) = \frac{P(A, B)}{P(B)} \implies P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{4.10}$$

When two nodes are directly connected (eg. nodes D and C in figure 4.2) then these two random variables directly influence each other regardless of any evidence about the values of the rest of the variables. In the case of an indirect connection one can distinguish four cases of variable interactions (Koller & Friedman 2009): Indirect causal effect, indirect evidential effect, common cause and common effect.

Indirect causal effect occurs when a path between two variables does not contain any variables for which evidence exists. In the example of figure 4.2, the nodes B and C interact with each other only if node D is not observed. Intuitively, one can think that if node D is observed, then information about the node B doesn't change the beliefs for the values of node C. Indirect evidential effect is

the symmetrical opposite of the previous case. Specifically, the values of node B are influenced by the outcomes of C only when D is not observed. This is due to the symmetrical notion of dependence, that is if $(B \perp C)$ then $(C \perp B)$.

A case of common cause occurs between nodes C, D, E in figure 4.2. In this configuration, information about C can change the levels of belief about D if D is not observed, and hence C and E are correlated. However, in the case that there exists evidence for the state of node D, then all information necessary to determine C is contained in D, and thus C and E are independent given D.

Common effect occurs in a configuration of the type B, D, A in figure 4.2. Contrary to the previous cases, observing node D activates the path between B and A. Intuitively, information about B and A is correlated when the outcome of D is observed, as both B and A are influenced by this outcome.

The above rules generalise for any number of nodes and paths in a DAG and guarantee the soundness and completeness of independencies in the graph (Koller & Friedman 2009). These rules are commonly referred to a directional separation or d-separation.

Similarly to UGM, the following sections will discuss different types of BNs in applications related to transportation and accessibility.

4.4.2.1 BNs in statistical inference

The conditional independence property of BNs is very powerful and can be applied in many different contexts. For statistical inference, BNs can naturally represent hierarchical dependencies where sharing of information or "pooling" occurs. Bayesian hierarchical models are statistical models that are formulated based on this property. Figure 4.3 shows a 2-level Bayesian hierarchical model where the observations (y_i) are assumed to be influenced by a set of parameters that are sampled from an underlying normal distribution with parameters μ, σ:

The hierarchical structure is achieved by placing prior information for the model parameters at different levels in the representation. The prior information is translated into the prior distribution which, combined with the likelihood function that represents the theoretical distribution of the data, forms the posterior distribution. In the example of the 2-way hierarchical model (figure 4.3) this is given by Eq: 4.11.

$$p(\mu, \sigma, \theta, \alpha|y) \propto p(y|\theta)p(\vartheta|\alpha)p(\alpha) \tag{4.11}$$

In terms of applications, Perrakis et al. (2012) used a Poisson-Gamma Bayesian hierarchical model to estimate OD flows from census data, as a viable alternative

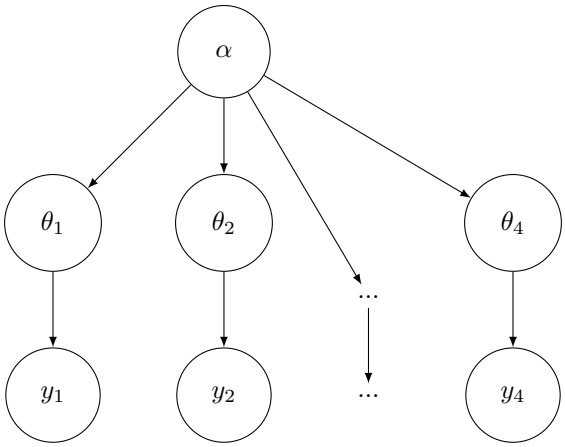


Figure 4.3: A 2-level Bayesian hierarchical model

to gravity type land use transport interaction models. Using the assumption that socio-economic variables scale over different aggregated levels, they used a regression framework to predict and generate origin destination flows from different geographical aggregations as well as assess the significance of the covariates in the predicted flows. In the context of travel behaviour modelling and discrete choice modelling in particular, Daziano et al. (2013) argued that the concept of subjective knowledge and preferences that is inherently present in choice models, can be efficiently represented by the prior distribution. As a case study, they used a Bayesian multinomial probit model to infer transportation mode choice from revealed preference data on interurban travel choice in Canada. They found that a carefully selected prior distributions not only can account for weakly identified parameter estimates but can help alleviate biases compared to the classical statistical (frequentist) approach, especially in the case of limited observations.

Despite the advantages of BNs (e.g. capturing non-linearities between variables, providing a solid mathematical framework for incorporating prior knowledge as well as representing the causal structure of a phenomenon), they have rarely been applied in the field of CA based applications, up until very recently. In particular, drawing from CA framework to measuring well-being, Ceriani et al. (2016) used the causal structure of BN to discover the flow of influence between variables related to well-being as measured by the European Bank of Development Life in Transition Survey. The information flow captured by the model revealed some interesting facts, notably the way objective life circumstances (such as age, employment status etc.) affect subjective beliefs about aspects of well-being (eg. personal voice and political views, social networks and ties) and the way these manifest in the outcome of questions related to life satisfaction.

Reasons for the slow adoption of BNs in CA studies could be found to the

disadvantages of BN such as the computational intensive framework of Bayesian statistics, combined with the predominantly frequentist statistical approach of social theory studies.

4.4.2.2 BNs in computational intelligence

Similarly to MRFs, BNs have also been used for knowledge discovery from raw mobility data. Relevant application domains include travel mode as well as activity purpose detection.

Exploring the applicability of BNs to travel mode detection from GPS signals, (Xiao et al. 2015) used the causal information flow between acceleration, speed, travel distance and travel heading to classify trip segments to walking, cycling, using the public transport or driving. Using data from 202 individuals, they were able to use the BN structure to capture the interrelationships between the variables as well as addressing uncertainties in measurement. Their model provided increased accuracy in travel mode detection compared to widely used machine learning algorithms for this task (such as support vector machines and artificial neural networks). Arguing that utility based accessibility methods are not in line with evidence on human behaviour (people are more likely to reason under what if scenarios rather than in terms of utility maximisation), Janssens et al. (2006) used a combination of a decision tree and a BN to account for interpretability of an individual's decision process to reach and perform an activity. Specifically, they used a BN to derive the skeleton of an individual's decision making process using travel diary data, before feeding the resulting rules into a decision tree for more direct interpretation. They found that this augmented approach can provide a more accurate modelling framework for a complete accessibility modelling framework compared to decision trees. However, comparing the same approach with a BN, the authors found insignificant differences in prediction accuracy.

Bayesian networks for computational intelligence applications can be naturally extended to incorporate the temporal dimension of mobility. Most commonly, this is done through the use of a markov chain that imposes a temporal dependence between a sequence of random variables. A frequently used dynamic BN for sequential data is a Hidden Markov Model (HMM). Figure 4.4 below shows the graphical structure of a HMM:

Within an HMM, the temporal dependency is represented by a stochastic latent process conditioned on the actual observations. Equation 4.12 shows the joint density of figure's 4.4 HMM:

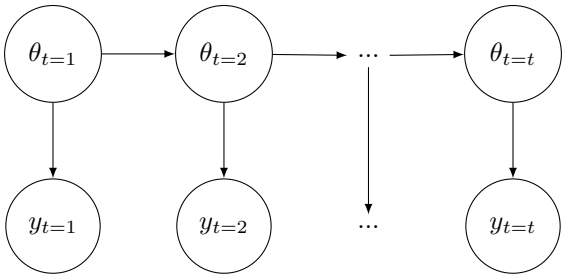


Figure 4.4: A simple Hidden Markov Model

$$p(\theta_1, \theta_2, ...\theta_t, y_1, y_2..., y_t)p(y_1|\theta_1)\prod_{t=2}^{t} p(\theta_t|\theta_{t-1})p(y_t|\theta_t) \tag{4.12}$$

In terms of applications, there is a wealth of studies on trajectory classification using dynamic BNs (Zheng 2015) for both activity detection and transportation mode choice from raw mobility data. For example, Bantis & Haworth (2017) used a dynamic BN to infer individual's mode choice from low accuracy smartphone data. Using information related to an individual's transportation behaviour from past travel surveys as prior information, they were able to assess the significance of individual characteristics to mode choice while at the same time classifying untagged location traces into travel modes. In the field of activity type detection, Bantis & Haworth (2019) used a dynamic BN model with social media data to benchmark the limits of accuracy of activity type detection from unlabelled data. More in depth analysis of both of these studies will be given in subsequent chapters.

In another study, Song et al. (2014) used the generative properties of a HMM to predict future locations of individuals from GPS mobility traces following an earthquake. Specifically, by using data obtained from smartphone GPS traces following the Great East Japan Earthquake and Fukushima nuclear accident together with disaster reports, they found that people behaviour tend to persist as normal at the early stages of the disaster. Significant deviations from the usual behaviour occurred hours/days after the event. Activity prediction was also the topic of Ye et al. (2013). Using check-in data from a social media app, the authors employed a two level HMM to infer the most likely activity category given an individual's past activities. Following from this, they were able to use location and time specific covariates to suggest the most likely location of the activity.

Finally, in terms of robustness and prediction accuracy for combined transportation mode and activity detection purposes from raw mobility data, Feng & Timmermans (2016) found that BNs outperform classifiers such as Decision Trees (DT), Support Vector Machines (SVM) and Logistic Regression (LR).

4.5 Structural Equation Models

Another type of models that draw on the benefits of representing relationships through a directed graphical structure isStructural Equation Models (SEM). Contrary to BNs where the modelling approach is inherently stochastic and the interactions are captured through conditional probabilities, SEMs use deterministic equations used to link variables with cause and effect relationships (Pearl 2009):

$$y_i = f(pa_i, \epsilon_i) \tag{4.13}$$

where pa_i are the parent nodes of y_i and ϵ_i is the error term.

The functions $f(pa_i, \epsilon_i)$ in SEMs are commonly derived from linear combinations of covariates $\sum \beta x_\kappa$, where x_κ can be defined as a parent node from another structural equation function. Contrary to BNs, SEMs graphs are allowed to form cyclic structures. Figure 4.5 shows a simple 2-way SEM:

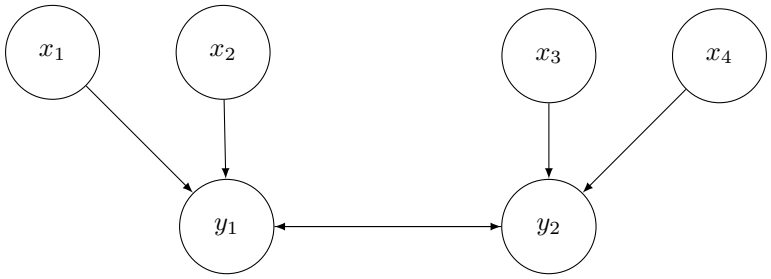


Figure 4.5: A 2-level Structural Equation Model

In this model, the structural equations are defined by the tuple:

$$\begin{aligned} y_1 &= w_1 y_2 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_1 \\ y_2 &= w_2 y_1 + \beta_3 x_3 + \beta_4 x_4 + \epsilon_2 \end{aligned} \tag{4.14}$$

In the field of transportation research, SEM models have been extensively used in travel behaviour/travel demand applications (Golob 2003). For example, Paulssen et al. (2014) used the hierarchical structure of a SEM to model the travel behaviour of individuals using the values-attitudes-behaviours model of cognition. Specifically, the authors used a questionnaire survey to gain insights related to attitudes in transportation such as flexibility and ownership, values such as power/hedonism and security and personal characteristics such as age, gender and income. Using a two-layer linear regression SEM as input to the non-stochastic component of a utility-based choice model, they found a significant influence of

personal characteristics to attitudes and values that contribute towards the choice of transportation mode. Investigating the interactions between travel behaviour, place-based accessibility and personal characteristics Simma & Axhausen (2003) used a SEM to assess the significance of endogenous variables (such as car ownership and gender) and exogenous variables (such as proximity to amenities) to transportation mode choice. They found that the influence of personal characteristics had stronger effect on individual's choice compared to place-based variables such as the number of reachable facilities.

The ability of SEMs to capture endogenous and exogenous covariates as well as interactions, led some authors to consider the use of SEMs as a way to operationalise the CA framework for different applications. Exploring the capabilities of children for being educated and being adequately sheltered, Krishnakumar & Ballon (2008) used a latent SEM to estimate the interrelations between the latent capabilities, endogenous and exogenous factors (such as household composition, personal characteristics, availability and accessibility of educational institutions). Their approach was based on a set of linear equations relating the different strata of the SEM through the covariance matrices of the structural equations coefficients. In another study, Di Tommaso (2007) used a variant of SEM, Multiple Indicator Multiple Causes model (MIMIC) to capture the significance of personal, household and environment characteristics on children's basic capabilities. Within a MIMIC model, the covariate variables are used as input to the latent component of the model and this relationship defines the structural part of the model. The measurement part of a MIMIC model links the latent component with the response components that constitute the measured variables.

4.6 Advantages and disadvantages between causal graphical models

The previous sections explored different classes of graphical models that could be used to model a CA accessibility framework. From the multitude of graphical model types that exists in the literature, only the ones that can be used to capture causal relationships in a hierarchical manner were reviewed. In this section, the advantages and disadvantages of these modelling frameworks relative to each other are assessed.

Structural equation models represent relationships between variables using algebraic equations (Bottou et al. 2013), which, in the majority of real world applications, are linear (Congdon 2007). This is achieved through a regression relationship between the structural equations, implemented through the use of covariance

matrices between the regression coefficients (referred to as loadings in SEM literature). One of the most important advantages of SEM is the model's ability to capture cyclic effects between variables. As such, reinforcing influence between a cause and effect relationship can be effectively represented. Another advantage mentioned in the literature is the ability of SEMs to represent counterfactuals (Pearl 2009). Counterfactuals are statements relating to outcomes that were not observed. For example, a counterfactual would be the statement: Would X have taken the bus if the bus stop was not crowded? In this case, the observed variables are that X did not take the bus and the bus stop was crowded. SEM captures this statement by means of three assumptions (Lucas & Kemp 2015): First, events that appear stochastic would be revealed as deterministic had all causal variables been known. Practically in SEMs, this is accounted by including latent exogenous variables. Second, these exogenous variables retain their values in both realised and non realised outcomes. The third assumption relates to interventions. An SEM assumes that a counterfactual outcome can occur only by a direct intervention. Returning to the example above, this corresponds to the question: Would taking the bus had occurred if we had forced the bus stop to be less full without influencing the causes of bus stop busyness in any way that would introduce side effects?

One disadvantage of SEMs is their limited ability to capture uncertainty in the parameter estimates. This is especially true in the case of variables omitted from a representation of phenomenon, either due to its complexity or due to the lack of available data. Although SEMs incorporate any unobserved effects in the random error term, this does not necessarily solve the problem of biased parameter estimates as typically researchers impose a structure on these errors (most commonly normality constraints) (Tomarken & Baker 2003). Furthermore, the assumptions of SEMs that relate to reasoning behind counterfactual arguments are open to criticism. According to (Lucas & Kemp 2015), it is natural to assume that people adjust their assumptions about a cause of a phenomenon in order to explain a counterfactual. Such type of inference is not possible in a SEM. Moreover, the premise of a counterfactual as a result of an intervention is in contrast with the counterfactual as if it was not-observed. Referring back to the example, people are more likely to readjust their assumptions if they observe a non crowded bus stop rather than if the bus stop was empty due to an intervention.

Both BNs and MRFs are stochastic models representing the relationship between random variables either through conditional probabilities or joint probabilities of their Markov blanket (factor potentials). It is important to note that these two classes of PGMs are not mutually exclusive as models involving products of

potentials and conditional probabilities do exist and are commonly encountered with hierarchical models (Buntine 1995). The inherent stochasticity of PGMs allow a sound representation of uncertainty for every random variable in the graph (Koller & Friedman 2009). Being generative models, PGMs provide a complete modelling framework that can be used to generate synthetic data given the model, thus enabling the exploration of what-if scenarios. Moreover, PGMs allow a more flexible representation on the nature of interactions between variables, as they don't assume linearity in the relationship between them.

The directed structure of BNs allows propagation of information through conditional probabilities, enabling thus causal inference. Being acyclic by definitions, BNs do not allow cycles in their structure as this would compromise the validity of the node distributions. As a result, they can not model cyclic cause and effect relationships. MRFs on the other hand, are more flexible in representing interactions between variables. However, it is more difficult to represent hierarchical cause/effect relationships due to the lack of directed edges in their graph representation.

BNs are insufficient for handling counterfactual arguments in their native form (Pearl 2009). This is because information contained in the conditional probabilities of the observed variables are not sufficient to uniquely determine countefactual arguments (Balke & Pearl 1994). However, the probabilistic structure of BNs allows easier integration of prior knowledge over the state of variables regardless the observations, which can lead to a fully Bayesian specification. This can be used to infer counterfactual arguments by sampling the posterior predictive distribution under counterfactual activity (Brodersen et al. 2015).

One very important advantage of BNs over SEMs and MRFs is their ability to effectively combine benefits related to inferring higher level semantic information from low level data, as well as assessing the significance of the relationships between the nodes of the model (see table 4.2). This is a very attractive property as it allows the use of raw mobility data within accessibility evaluations, as well as allowing the representation of accessibility measurement through the causal structure of the CA.

The table below summarises a selected literature related to applications of graphical models for different transportation related challenges.

Authors	**Graphical Model Type**	**Model Type**	**Data / Goal**	**Advantages / Disadvantages**

Xiao et al. (2015)	BN	BN	GPS traces / Travel mode detection	Good detection accuracy compared to other classification methods / No inclusion on characteristics of travelers or characteristics of environment
Widhalm et al. (2015)	UGM	RMN	CDR data / Activity detection	Model accounted for uncertainty due to low data precision / accuracy for both mobility data and land use types / No individual characteristics taken into account, model resulted in activity 'clusters' rather than recognising activity types
Perrakis et al. (2015)	BN	BHM	OD matrices / Trip distribution	Model incorporates both trip attraction and production in a probabilistic framework / Coarse modelling scale, no modal split, no attraction covariates
Sun et al. (2014)	BN	DP	GPS , accelerometers / Activity detection, transportation mode	Model agnostic to different input; no training data required / Datasets used represent ideal conditions

Golob (2000)	SEM	SEM	Travel diaries / Travel behaviour	Sociodemographic and external covariates are taken into account / Model doesn't take into consideration the spatial/temporal dependency between activities
Xiong et al. (2015)	BN	HMM	Travel diaries / Travel mode behaviour	Model captures travel mode choice evolution dynamics together with household characteristics / Transition dynamics are evaluated at an aggregated level - all transportation modes are assumed to be available to the users
Allahviranloo & Recker (2015)	UGM	CRMF	Geocoded travel diaries (as a surrogate for GPS) / Activity detection	Methodology combined the hierarchical nature of activity classification features with socioeconomic ans spatiotemporal characteristics / Ambiguous activity classes, moderate activity predictions

Golob & McNally (1997)	SEM	SEM	Travel diaries / Explaining household activity and demand for travel	Breakdown of travel demand for activities by gender / Assumptions of normality between travel times and activity duration hard to defend

Table 4.2: Literature summary

Finally, it is worth mentioning that in practice, graphical models have more similarities than differences (Pearl 2009). For example, although BNs represent a joint density over all random variables, the density factorises to individual quantities (the conditional probabilities and, in the case of Bayesian specification, prior probabilities). These quantities are thought to be equivalent to the structural equations by some authors (eg. in the case of stochastic regression $N(\beta_0+\beta_1 x, \sigma)$) (Bottou et al. 2013). This is also supported by a growing literature that implements both types of models as complimentary across different application domains (Duarte et al. 2011). In any case, for practical applications, a variety of approaches within a single modelling framework seem to be more appropriate. This means that the modelling approach should be able to incorporate both stochastic and deterministic variables as well as directed and undirected relationships depending on the modelling goal and the nature of the data.

4.7 Chapter summary

This chapter introduced the theory behind graphical models, following by a description of the most commonly used models of this class in transportation literature, with special focus on studies dealing with elements that could be associated with inference of different components associated with accessibility measurement.

From the different graphical model types, only BNs satisfy the requirements posed by the research objectives mentioned in chapter 1. These are summarised by the requirement of expressing accessibility through the causal structure of CA at an individual level, using unlabelled mobility data of different spatial and temporal resolution. These will be explored in the following chapters from two angles: chapter 6 section 6.2 will assess the ability of BNs to infer the mobility component of disabled individuals while assessing the degree of contribution of personal and environmental characteristics, while chapter 6 section 6.3 will assess the capability

of BNs to impute activity types from unlabelled data, a concept central in accessibility evaluations. Using the results of these case studies, an explicit formulation of accessibility through a CA will be given in chapter 6 section 6.4 through a case study using London's Automatic Fare Collection data.

Before introducing the methodology of this thesis, a description of the datasets used and the preprocessing steps followed will be given in chapter 5.

Chapter 5

Data description and preprocessing steps

5.1 Chapter overview

This chapter introduces the mobility data used in the case studies presented in the remainder of the thesis. In particular, section 6.2 uses mobility data generated from low resolution smartphone devices and data generated from conventional travel surveys in the context of exploring the relationship between personal characteristics, environment variables and individual mobility. In this way, the potential of using Dynamic Bayesian Networks (DBN) within the context of combining different factors and data sources is established. Section 6.3 uses a DBN together with low resolution online geo-location data to investigate the limits of activity type inference accuracy from unlabelled data. This knowledge is essential for the purposes of chapter 7 which uses a combination of unlabelled passive mobility data (AFC) and travel surveys (LTDS, RODS) to formulate the capabilities approach to accessibility model used to explore the link between social exclusion and transport disadvantage in London.

5.2 Data requirements

Approaching accessibility through the structure provided by the CA, requires a modelling framework capable of expressing accessibility components and their interactions in a hierarchical, structured way that enables statistical reasoning. To this extent, the assessment of the degree that these requirements are captured using DBNs would be beneficial prior to the case study of chapter 7. Furthermore, the premise of using a modelling framework that can use both travel surveys and passive mobility data at the level of an individual should be evaluated in a ro-

bust way using ground truth information, before scaling the methods to a fully unlabelled dataset. Section 5.3 describes the dataset collected for this purpose.

Validation of the methodological framework using ground truth data is particularly relevant in the context of activity type inference from unlabelled mobility data. Knowledge of activity types performed at a destination is out of reach from opportunistic datasets collected from service providers (such as AFC data), however they are one of the most important elements of accessibility evaluations. Therefore there is a need to assess the limits of activity type imputation accuracy using a dataset that provides ground truth information while at the same time mimicking the settings encountered with low resolution mobility data. The characteristics of the social media dataset described in section 5.4 makes it ideal for such purpose.

Finally, the potential of unlabelled passive mobility data for the purposes of accessibility evaluations within a recurring, cost effective and near real-time framework is explored using London's AFC dataset. Compared to traditional travel surveys, this dataset has some important properties that allows a deeper investigation of equality levels in accessibility. First, it consists of trajectory data at the level of the individual, allowing a more complete evaluation of mobility habits. Second, it is generated as part of Transport for London (TfL) day to day business, which makes for a continuous and cost effective update of individual trajectories. Third, it can be coupled with additional travel survey data through a unique passenger ID, allowing investigation of the influence of socioeconomic characteristics on the levels of equality experienced by individuals. This dataset is further described in section 5.5.

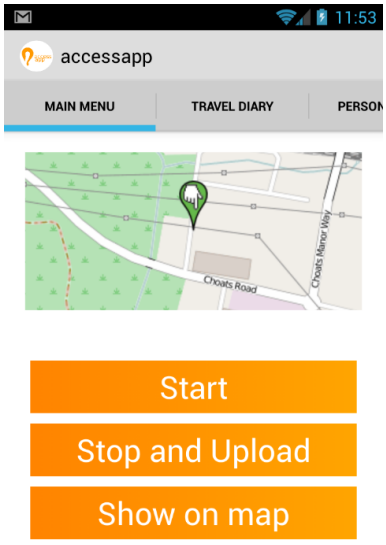
5.3 Low resolution smartphone data

For this mobility dataset, data from a bespoke developed geo-enabled smart-phone application was used. The use of smart-phone applications in mobility studies has seen increased interest from researchers in recent years in the fields of activity and transportation mode detection (Montini et al. 2015, Kim et al. 2014, Widhalm et al. 2012). The main advantages of using such an approach are the ease and cost efficiency of data collection and the potential to achieve high spatial accuracy. The latter is due to the prevalence of smart-phones equipped with GPS receivers and accelerometers. On the other hand, the most common disadvantage that has been reported is the increased battery demand on user's devices (Xiao et al. 2015, Wu et al. 2016), especially when both GPS and accelerometer readings are logged.

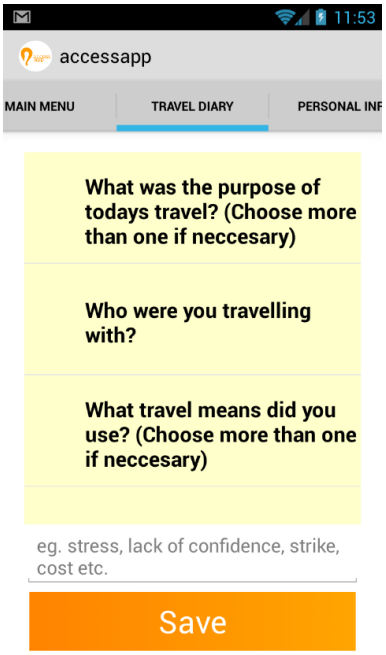
However, for any practical applications, the accuracy and precision of the

collected data is low, and this imposes an additional challenge for modelling (Eftekhari & Ghatee 2016, Wu et al. 2016). This can be especially true for middle to low end smart-phones.

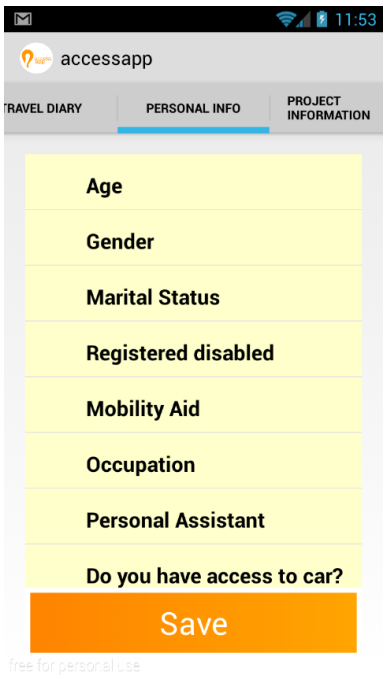
Figure 5.1 shows some snapshots of the developed application.



(a) Main menu of the app



(b) Travel diary section



(c) Personal information section

(d) General information section

Figure 5.1: Accessapp screenshots

5.3.1 Data collection process and data idiosyncrasies

The developed smartphone application makes use of android and iOS location Application Programming Interfaces (APIs) to log a user's coordinates in an non-intrusive way, while simultaneously managing the trade-off between battery use and coordinate logging[1][2].

The app was given to two individuals experiencing mobility difficulties. The first volunteer is 40-59 years old, female, full-time employed and a crutches user while the second is a 22-39 year old male, full-time employed and a mobility scooter user. Accessapp was configured to record coordinates at regular time intervals (2 minutes) if there is a significant distance between two subsequent position fixes. This distance was taken to be 30 meters. All information was stored in a text file in the internal storage of the phones which was then uploaded to a secure FTP (File Transfer Protocol) server located at UCL computer science department. All information was anonymised by a unique user ID during storing and uploading. The data collection, storing and processing aspects of experiment has been approved by UCL ethics committee (project ID 7111/001, see appendix F).

The raw locations are presented in figure 5.2 below. For the crutches user, the temporal window of observations was 7 days in total, while for the wheelchair user it was 3 days in total.

(a) Crutches user

(b) Wheelchair user

Figure 5.2: Raw mobile phone location data

The main disadvantage of pipelining the location logging process through the use of an API is that the researcher has limited control over location accuracy and temporal resolution. As the API uses different sensors to determine an individual's location, the more sensors it uses the greater the location accuracy. Depending on factors such as sensor availability at the moment of update and battery level, this can vary. For example, a fused GPS/Wi-Fi update can have accuracy in the

[1] https://developers.google.com/maps/documentation/geolocation/intro
[2] https://developer.apple.com/documentation/corelocation

order of tens of meters, while a GSM cell-tower update can have accuracy in the order of hundreds of metres.

Moreover, the resulting datasets are characterised by data points with inconsistent temporal resolution, as the sampling interval is determined by the background smartphone applications that make use of location services. Figure 5.3 below shows the relationship between spatial accuracy and temporal resolution for the two participants. The spatial accuracy was obtained by querying the API for the estimated confidence interval of the location estimate.

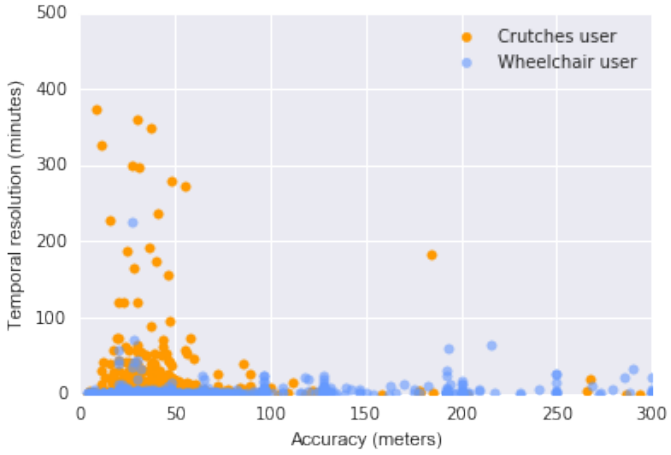
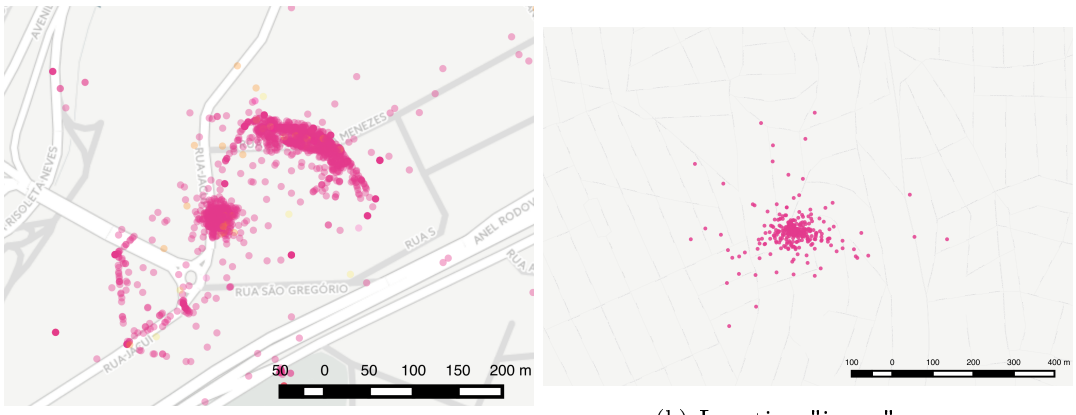


Figure 5.3: Accuracy vs resolution for the two participants

Such artefacts can have an important effect on the overall classification task by adding systematic (such as location "drift") and non systematic noise (such as sudden "jumps" in location) that influence the regularity of point patterns, thus increasing variability of the classification features, such as speed (Figures 5.4, 5.5). This adds increased ambiguity in distinguishing the transportation modes used, increasing the overlap between mobility states, especially for modes that are characterised by more subtle changes, such as walking or dwelling.



(a) Location "drift"

(b) Location "jump"

Figure 5.4: Examples of “jump” and “drift”

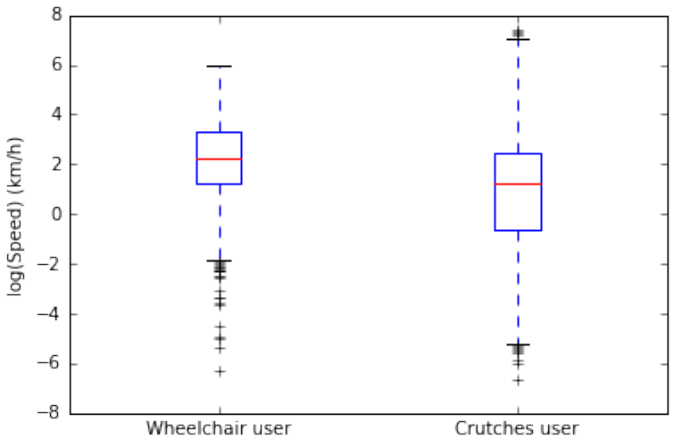


Figure 5.5: Boxplots showing the variability of the classification feature. For visualisation purposes, the speed was log transformed.

5.4 Low resolution online geo-location data

For the purposes of activity type inference, the mobility data used were obtained from the location-based social network Foursquare for the Greater London Area. Foursquare is a search-and-discovery location based service for smartphone users that allows sharing of visited places via the check-in option. The service was created in 2008 and was initially designed as a game. However, it very soon evolved into a large scale social network community serving as a recommendation engine around physical places (Noulas et al. 2011). The development of a dedicated and easy to use API allowed researchers to source Foursquare check-in data for many different research goals, ranging from activity discovery (Noulas et al. 2011) to prediction (Ye et al. 2013) to pattern classification (Hasan & Ukkusuri 2014). The choice of this dataset in the context of this study can be justified on the following premises:

- The sequential nature of individual check-ins (i.e publicising one's current location to the social network) can be regarded as a trajectory if the individual check-ins are connected chronologically (Zheng 2015). In this regard, it has many similarities with other mobility datasets that are characterised by chronologically ordered pairs of coordinates generated by a moving individual (eg. AFC, CDR, location based social networks).

- Foursquare check-in data are associated with an individual's disclosure of location together with semantic information on the nature of the location (eg. restaurant, university etc.). The disclosed activity types can serve as ground truth dataset to test the accuracy of the activity inference algorithm.

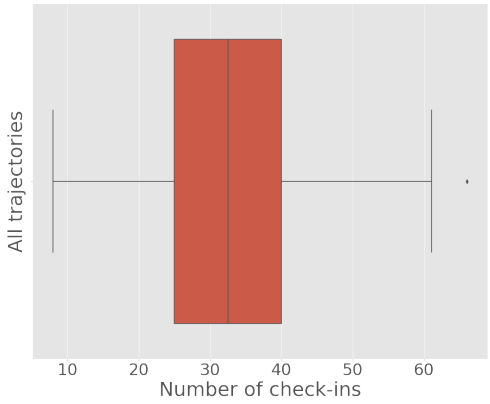
- Foursquare data holds and maintains a comprehensive database of POIs that can be used in conjunction with the the check-in data for activity inference.

- As Foursquare data are primarily focused around leisure/entertainment activities, they can be used to explore an individual's non-commuting activity patterns.

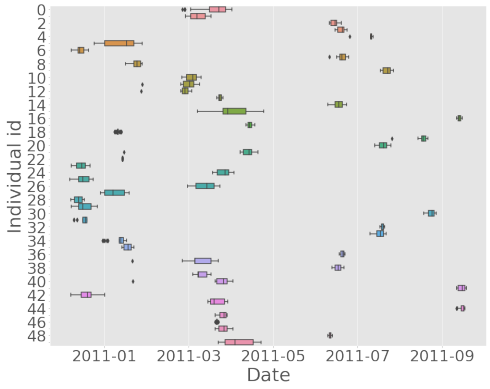
Using the Foursquare API, check-in data along with venue information were sourced for a period of 10 months (2010/12/31 - 2011/09/30). The following sections describe the data preprocessing steps followed to reach to the effective sample size used for this study.

5.4.1 Data preprocessing

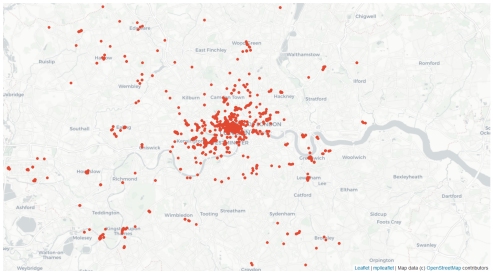
The vast majority of sourced Foursquare check-in data contain infrequent users that use the service occasionally. For the purposes of this thesis, it is important that a trajectory dataset for each individual is obtained in a way that resembles other unlabelled mobility datasets (such as AFC or CDR data), so that the methods and findings can be transferred to the case study using the AFC dataset (chapter 6 section 6.4). For this reason, individuals that used the service with interruptions between consecutive check-ins of more than a week were not included in the analysis. This resulted in a decrease of the overall sample size as well as the spatial and temporal extent of the study which should be taken into consideration when interpreting the outputs. On the other hand, this step removes the bias in activity type imputation accuracy that may result from applying the methodological framework to trajectories that are significantly different from the AFC dataset used in chapter 6 section 6.4. The term "trajectory" in the context of this study is the set of all check-ins for each individual Foursquare user throughout the study period. The resulting dataset contained 50 unique users. The average number of check-ins in each trajectory was 33, while the average time span for those was 11 days (note that each day can contain multiple check-ins). Figure 5.6 below shows the distribution of check-ins counts for all individuals in the sample, the date range of check-ins for each individual, as well as the spatial distribution of check-ins.



(a) Boxplot of check-ins for all trajectories.



(b) Boxplots of time span of check-ins for each individual trajectory.



(c) Spatial distribution of check-ins for all trajectories.

Figure 5.6: Foursquare POI and labour demand data along with the downsampled dataset

5.4.2 Activity detection feature space

Similar to past literature on activity type inference using mobility data (Chen et al. 2016), a Points of Interest (POI) database was used as the activity detection feature space. Specific approaches on the way POIs are used for the task vary in the literature. Huang et al. (2010) introduced the notion of a geometric construct for each POI that is a function of static parameters, such as POI footprint, attractiveness (popularity of the POI) as well as temporal parameters such as time of day and day of the week. They then evaluated the intersections of an individual's GPS trajectory with respect to this construct to determine the activity of an individual, with the highest number of potential intersections determining whether the POI is selected as an activity place. In another study, Yuan et al. (2012) assigned a POI vector to regions in the city derived from the road network geometry. Together with GPS data, they used this vector within an Latent Dirichlet Allocation (LDA) model to assign a function to each region. A similar model was used by

Zhang et al. (2016) in the context of discovering common interests from individual trajectories. Within the context of LDA, the authors used a POI database as an analogy to words in topic modelling. The POI vector that corresponded to the topic to be discovered was defined using a buffer area around bus stops with the underlying POI database.

Determination of the bounding area of the POI feature vector was driven by two factors:

- The need for a generalisation of the methodology to other mobility datasets as far as possible.
- The need to establish an accuracy assessment framework under different configurations of POI feature vectors.

Specifically, the activity feature space is defined to be the area bounded by different walking isochrone levels (levels of equal walking time) centered at a Foursquare check-in location on the road network. An isochrone based approach is common in transport planning and accessibility studies (Transport for London 2010, Dodson et al. 2006, Wu & Hine 2003). Moreover, as already mentioned in section 2.2.1.2, an isochrone approach is often applied within the context of cumulative-based accessibility indicators, accounting for the potential destinations that could be reached from the check-in location. As mentioned in section 3.4, this potential is of major importance for representing capabilities.

The isochrone levels were chosen to reflect different accuracy levels of mobility data, corresponding to walking distance along the road network ranging from 5 to 20 minutes at 5 minute intervals. For the computation, data from the open source road network database OpenStreetMap was used, assuming a constant walking speed of 4.5 km/hr.

This process generates areas that can be related to mobility data of varying precision. For example, the case of the 10 minute level isochrone corresponds to an approximate distance from the check-in point of 300-500 meters, precision often encountered with mobile phone cell-tower data, depending on the antenna configuration (Widhalm et al. 2015).

Next, the individual Foursquare POI venue names were aggregated to higher level categories using the default Foursquare category hierarchy to ensure consistency between the low level POIs and the higher level activity types. This includes categories such as "Arts and Entertainment", "Colleges and Universities", "Food", "Outdoors and Recreation" and "Shops and Services". Finally, the POI feature vector was defined to be the POI counts per individual category that intersect

each area bounded by the isochrones. The distribution of POIs across the study region is shown in figure 5.7.

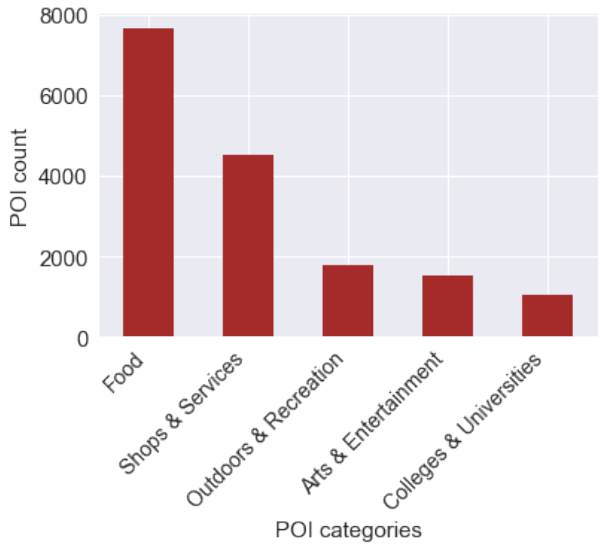
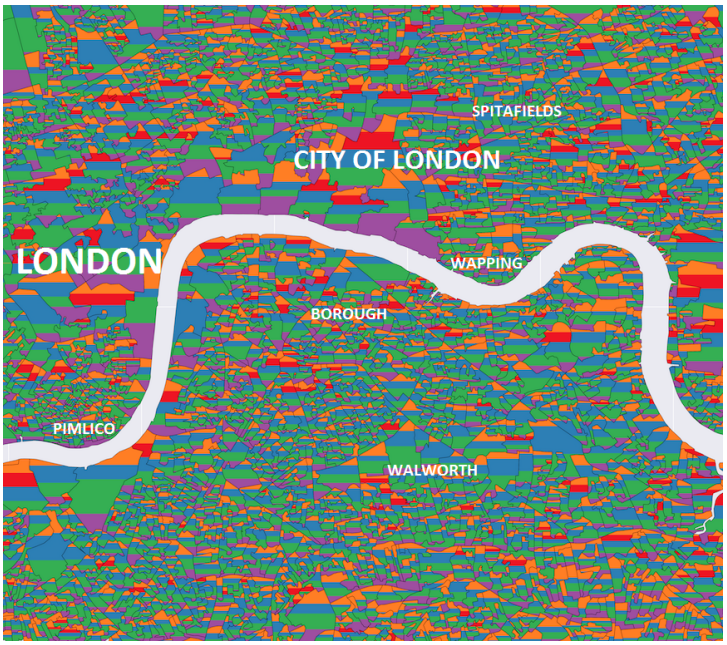


Figure 5.7: Distribution of POIs throughout the study area.

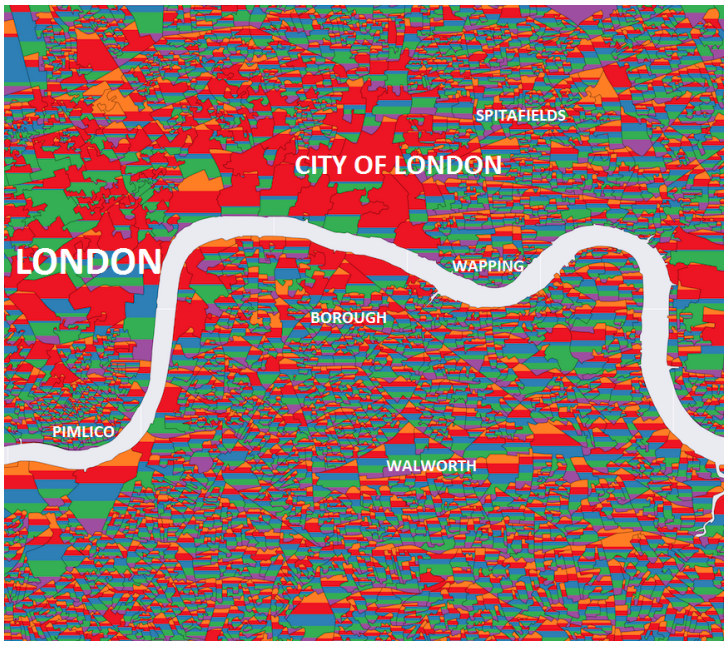
Figure 5.7 suggests a considerable imbalance among the POI categories that reflects the core purpose of Foursquare service: allowing users to share their leisure/entertainment activities. Imbalanced datasets have been the subject of a significant body of research as they can deteriorate the performance of any classification/clustering algorithm by introducing a bias towards the majority class (Krawczyk 2016). As a result, there have been numerous attempts to alleviate this problem ranging from simple random under/oversampling of the majority/minority class to more sophisticated ones exploiting the structure of the classification feature space (eg. ADASYN, SMOTE). Within an unsupervised classification setting, the problem of an imbalanced dataset becomes even more complicated as there is no training set to assist in the identification of the minority class in order to equalise the dataset accordingly. In light of this, this study used land use information to downsample the Foursquare POI vector within each activity detection isochrone polygon. The degree of undersampling for each activity class was calculated using the UK's 2011 Census labour demand data as the fraction of the total count of jobs in each isochrone polygon. This includes counts of jobs for 20 industry sectors at an 'output area' geographic aggregation level. This geography corresponds to polygons that are adjusted to contain at least 40 households, the target size being 125 households. For this study, 4 industry sectors were used in line with the Foursquare POI categories: Education; Wholesale and retail trade; Accommodation and food service activities; Arts, entertainment and

recreation. For the activity category 'Outdoors & Recreation' the ratio of green spaces to the general output area was used, as derived from the OpenStreetMap land cover dataset. The final dataset is the product of an elementwise multiplication of two vectors for each isochrone polygon: the original Foursquare POI vector and the vector containing the fraction of jobs to the total number of jobs per activity class, as derived from the labour demand data.

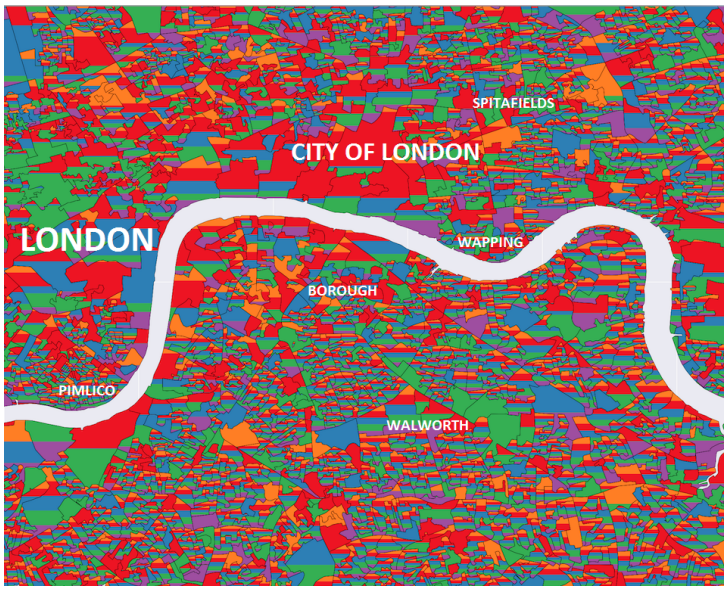
For illustration purposes, the Figure 5.8a displays the proportion of activity categories using the labour demand data within each output area, Figure 5.8b shows the resulting proportion of Foursquare POIs per activity (displayed as aggregated counts per output area), and the final proportion of POIs after applying undersampling is shown in Figure 5.8c. The black bounded polygons in this figure represent the extent of the OA while the size of the individual colored patches inside each OA correspond to the ratio of each activity category with respect to the sum of all activities inside the OA.



(a) Ratio of labour demand data categories per OA.



(b) Ratio of Foursquare POI categories per OA.



(c) Ratio of Foursquare POIs after performing downsampling. (d)

Figure 5.8: Foursquare POI and labour demand data along with the downsampled dataset

As it can be seen in Figure 5.8c, the final dataset maintains the general shape of the spatial distribution of Foursquare POI activity types (Figure 5.8b), while at the same time allowing for additional clusters to form (e.g. the Elephant and Castle shopping area around Walworth), a result of the undersampling process using the labour demand dataset (Figure 5.8a).

It is important to note that the resulting dataset is treated in a completely unsupervised setting, using only mobility characteristics of each trajectory and the isochrone derived POI vector in the model.

5.5 Oyster card/London Travel Demand Survey data

This section will begin by providing a description of London's AFC system, the Oyster Card dataset, before proceeding in explaining the association between the Oyster card and London Travel Demand Survey (LTDS) data. Then, it will describe the sequence of preprocessing steps undertaken to produce the dataset that is used in the case study of chapter 7.

5.5.1 Automatic Fare Collection Systems

Automatic Fare Collection Systems (AFC) systems were introduced as an alternative to traditional ticketing services, completely replacing or supplementing paper tickets with RF-ID (Radio Frequency ID) based reusable cards for some cities (Blythe 2004). As a technology, AFC is not new and has been used by transport service providers for almost 20 years. AFC offers a structured way of collecting financial and trip data of passengers and, in many cases, personal information such as age, gender and disability status. Transport service providers use this information to manage fare collection, help relieve passengers from some of the burden of manual ticket validation and improve security and overall user experience.

Besides managing the transport service, this structured way of representing passenger journeys has opened up a range of opportunities for applications that range beyond the original scope of the technology. Specifically, AFC data allows researchers to explore issues related to service reliability (Uniman et al. 2010, Freemark 2013, Wang et al. 2011), demand forecasting (eg. reconstruction of origin destination matrices) (Zhao et al. 2007, Barry et al. 2002), investigating human mobility patterns (Foell et al. 2014), applying potential accessibility measures (Smith, Quercia & Capra 2012) as well as inferring trip destination types (Han & Sohn 2016).

Compared to conventional interviews/travel diary studies, AFC data provide several advantages (Pelletier et al. 2011, Bagchi & White 2005). These range from

practical advantages such as reducing the burden on users as well as reducing the cost of data collection process, to advantages relating to the nature of the sample such as offering larger sample size for different population groups (if AFC data are linked to socioeconomic characteristics). Moreover, AFC data allow access to continuous trip data over longer periods of time, thus enabling longitudinal studies.

On the other hand, such data present additional modelling challenges to researchers. These can be summarised to lack of labels, sparseness, low spatial and temporal resolution as well as the lack of validation/reference datasets for activity types performed at destinations. Furthermore, is important to note that machine generated data such as AFC data, are only relevant to the individuals using the services that generated the data.

The problem of inferring activity types from AFC data is closely related to the problem of extracting semantic information from unlabelled mobility data discussed in section 2.3.2. To summarise, activity imputation from AFC data is a complex problem that is actively pursued using a very diverse set of methodological frameworks, ranging from simple rule based approaches, to discrete choice analysis and network/spatial statistical methods, to advanced machine learning and probabilistic methods. In terms of secondary data used to assist inference, a wide range of different sources have been utilised, including POIs, land use information, data from social media applications and household surveys.

The implications of the third challenge mentioned above are wider and are directly related to the scope of the studies using AFC data from public transport service providers: Inference/analysis results from AFC data are only relevant in the context of public transport use by an individual. This limits the scope of accessibility analysis as, by definition, such data do not cover journeys made by cycling or walking, journeys made by complementary modes of travel (such as dial a ride services), journeys made by private transport modes (car, taxis) etc. This is of particular importance and must be kept in mind by policy makers when interpreting modelling results derived from such data.

Nevertheless, the benefits of AFC data, especially the potential for longitudinal analysis at the individual level, make them an attractive complimentary data source to traditional accessibility audits. This is especially true within urban settings where public transport accounts for nearly 35% of the total journeys made (Transport for London 2011).

5.5.2 Data description

This section provides a description of London's Oyster card AFC system, along with secondary data that are used: the London Travel Demand Survey and Ordnance Survey POI data.

5.5.2.1 Oyster card AFC

TfL's own AFC system uses RF-ID stamped cards (called Oyster cards) as a unified transportation ticketing system for many public means of travel. This includes the underground (including Overground service), National Rail and other rail services as well as buses and trams. Within these cards, information related to individual trips is captured each time the Oyster card is used. For rail services (including the underground) a passenger is required to "tap" their card on the Oyster card reader at the station at the beginning of the journey, at intermediate stops in case of changing to an overground service, and at the end of the journey during exit. The total amount of fare is then deducted from the Oyster card balance according to a zonal fare system. Users may also use other contactless payment types such as credit/debit cards or smartphones in place of Oyster cards, which are also recorded by TfL. Here, we use the term Oyster card to refer to all of these payment types.

Bus services use a different approach for fare collection using Oyster cards. As London buses implement a fixed fare approach, bus passengers are required to touch their Oyster card while boarding. A fixed amount of fare is then deducted from their Oyster card balance regardless of the alighting stop. As a consequence, bus records lack alighting information. The procedure to infer bus boarding stops and alighting stops is elaborated on in section 5.5.3

A brief description of the most important Oyster card dataset characteristics for this research is given in (Reades 2014):

- Dataset contains population groups, as determined by the different fare types (adult, children, student, elderly people, disabled people).
- Dataset contains enter/exit information in case of travel by rail, bus route number in case of travel by bus.
- Dataset contains transaction time/day information.
- Dataset contains unique pseudo ID (in the sense that the data remain anonymous) for each record, generated by TfL.

Recognising the limitations of this type of data architecture, TfL have sought to utilise ancillary datasets to overcome the lack of boarding information. One such dataset is London's Automatic Vehicle Location (AVL) system, called iBus.

iBus is a collection of systems that enable real time location tracking and monitoring of London's bus fleet. Those systems include, among others, telematic technologies, tacheometers, GPS, gyroscopes and speedometers installed on every bus. iBus's ultimate goal is to record a bus's actions near stops. The way this is achieved is by recording four time stamps, each one signifying a specific bus operation, namely: bus is near a stop, bus is opening the doors, bus is closing the doors and bus is pulling away from the stop. These four records are then used to determine the approximate time that a bus is at a stop. Using this time-stamped information, TfL has developed a data matching algorithm that associates a particular AFC record with the corresponding bus stop as determined by the iBus data. Although the AFC dataset provided had already undergone the above described procedure to infer boarding stops, this was not the case for alighting bus stops and exit stations for tram.

The Oyster card dataset provided for this study is an 8 week sample from the end of October to the middle of December 2013, which TfL has prepared and shared with academic institutions for research purposes. Cleaning and canonicalisation of the dataset was done following Reades (2014).

5.5.2.2 London Travel Demand Survey (LTDS)

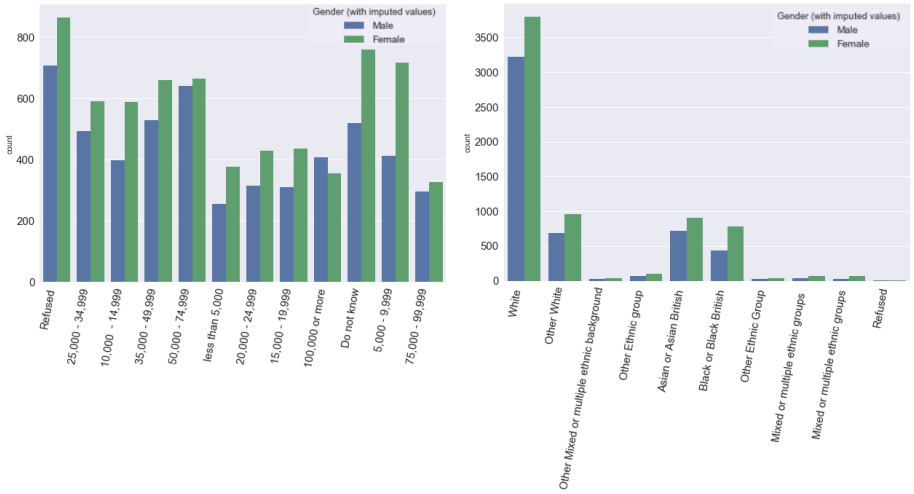
The Oyster card data does not contain personal information on sociodemographic characteristics. However, such information can be extracted from secondary data, in particular the LTDS.

The LTDS is an annual recurring questionnaire survey carried out at a household level aimed at probing TfL's customers' sociodemographic background and travel patterns, with a geographic coverage extending up to outer Greater London, within the M25 boundary. According to Transport for London (2011), during the survey all members of the household sample are interviewed and complete details of the travelling habits of the interviewees are recorded. The survey is comprised of three questionnaires: a household level questionnaire that collects socioeconomic and demographic details, an individual questionnaire capturing information on characteristics such as sex, age and health status, and finally a travel diary taken on the same day of the survey. The travel diary includes information about the travel mode and locations of origins and destinations.

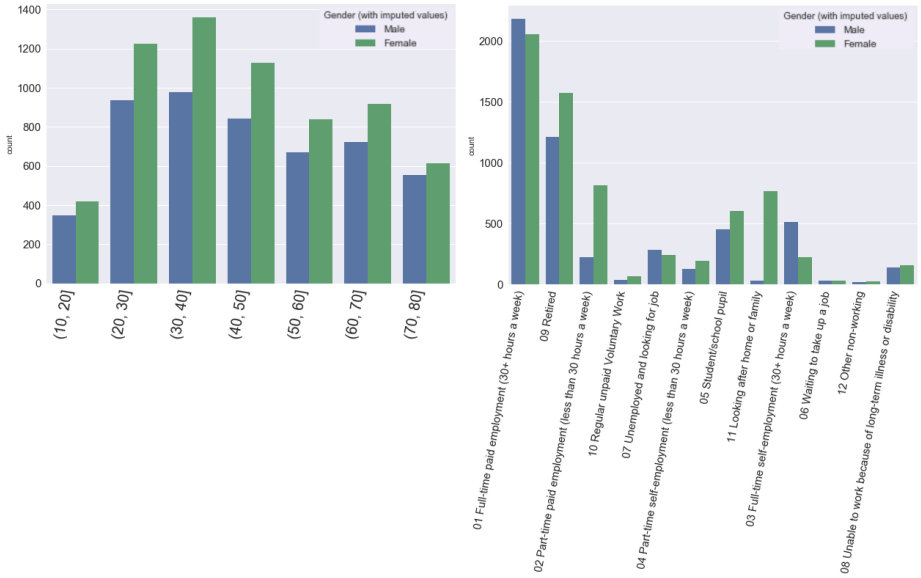
The sample size of the survey is approximately 20,000 individuals. The information is used by TfL to generate travel patterns which are used to improve its

services (Ortega-Tong 2013) and further understand the travel needs of London's citizens, particularly disadvantaged population groups (TfL 2014). Besides this, LTDS data have been used in research in a variety of different contexts. Exploring taxi drivers routing choices, Manley (2016) have used LTDS data to estimate the propensity of trip generation for specific activities in the context of a spatial interaction model. In another study, LTDS data have been used to correlate walking behaviours of young children based on household socioeconomic and environmental variables (Steinbach et al. 2012). Within a similar research objective, Sarkar et al. (2015) have used LTDS data to relate walkability of streets to the amount of green areas. The data-set provides very important insights linking mobility and socio-demographic characteristics for marginalised population groups. However, it lacks specific location information on the daily mobility habits of individuals that could be used to better inform transportation planning.

Figure 5.9 below shows some basic sociodemographic characteristics of the 2011/12 LTDS data.



(a) Income (left) and ethnic background (right)



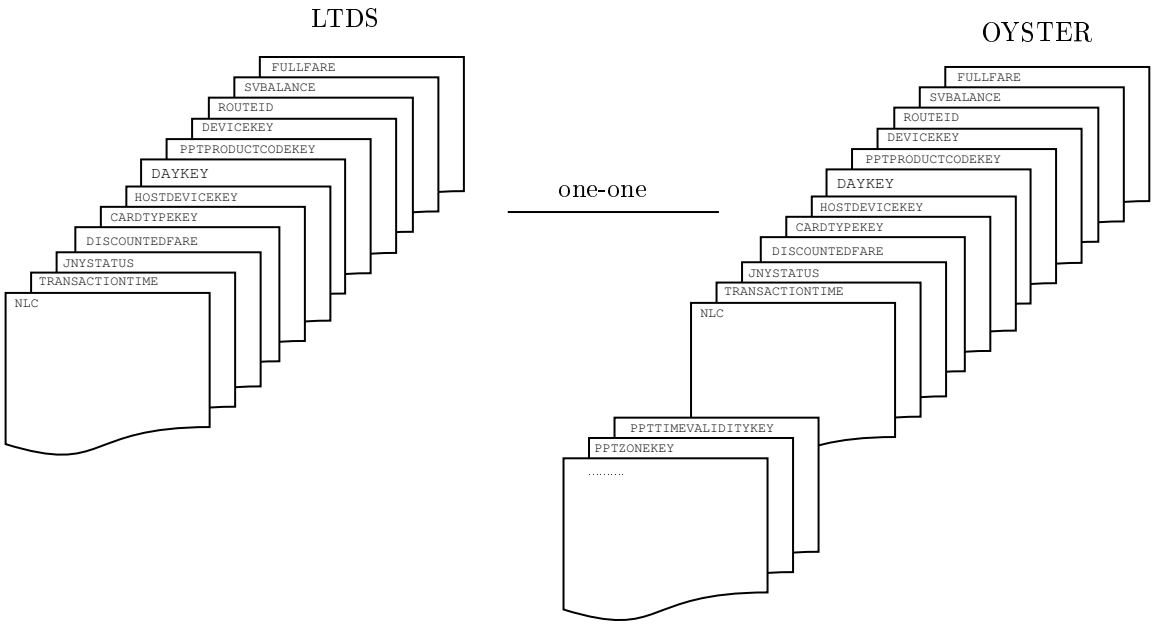
(b) Age group (left) and employment status (right)

Figure 5.9: General sociodemographic background of LTDS grouped by gender

During the 2011/2012 LTDS survey, respondents were asked if they were willing to provide their Oyster card unique ID for TfL to undertake further analysis of their travels. Since then, the relevant data has been stored by TfL's Customer Experience department, resulting in a database of approximately 12,000 cards and 10 million transactions from mid-June 2011 to March 2014.

Similar to iBus/Oyster card, a sample of this database was provided to academic institutions for research purposes. The time window of the data overlapped with that of the iBus/Oyster card sample described in section 5.5.2 for the period of October/mid-December 2013. However, the provided LTDS/Oyster database

```
 list IDS;
for i in LTDS unique PRESTIGEID do:
  for j in Oyster unique PRESTIGEID do:
     if exists:
```




```
  append j in IDS;
```

Figure 5.10: LTDS/Oyster matching process

sample contained only a subset of the original Oyster card column span and most importantly, it lacked the association of the raw iBus/Oyster card dataset with bus boarding stop. As bus boarding stop information is important for this case study, a matching process was necessary to reconstruct the individual trajectories for bus journeys and associate the raw iBus/Oyster card records with LTDS sociodemographic characteristics. This consisted of a one-to-one matching relationship between the two datasets for all columns of LTDS/Oyster database and the corresponding subset of iBus/Oyster card columns for each unique user ID (Figure 5.10).

Table 5.1 provides a description of the LTDS Oyster card columns that where used to match the iBus Oyster card records.

Table 5.1: Description of columns of Figure 5.10

Column	Description
DAYKEY	TfL day code
PPTPRODUCTCODEKEY	If a season ticket (as opposed to pay-as-you-go) of some type was in effect.
DEVICEKEY	Key of different Oyster card readers (eg. Gate, on bus, validation).
ROUTEID	The route id in case of a bus journey.
SVBALANCE	Balance amount.
FULLFARE	Indication of a full fare.
DISCOUNTEDFARE	Indication of a discounted fare.
CARDTYPEKEY	Key for different Oyster cards (eg. discount, elderly, staff etc.).
HOSTDEVICEKEY	Key for the device host.
NLC	National Location Code. A four-digit number allocated to every railway station and ticket issuing point.
TRANSACTIONTIME	Transaction time in minutes after midnight.
JNYSTATUS	Journey status (eg. Entry, exit, continuation, bus).

The matching process resulted in 224 unique Oyster card users. This is around 2.4% of the LTDS/Oyster database sample. This proportion might seem small, however, considering the limited iBus/Oyster card time window and the conservative process of the matching algorithm (not considering IDs from matching that contained corrupted records, partially matched records, incompatible length of matching records etc.), it is sufficient for the needs of chapter 7. In the absence of a complete matched dataset, this conservative data matching process minimises the risk of including erroneous samples in the case study which could introduce bias in the interpretation of the findings. This dataset will be referred to as *Oyster card/LTDS* for the rest of the thesis.

For the the case study of chapter 7, three population groups were defined, a group with low income individuals, individuals > 60 years old, and an unconstrained population sample (further details are given in chapter 7).

Figure 5.11 below shows the geographic distribution of visited places for each population group.

As can be seen, the majority of visited locations for the unconstrained and

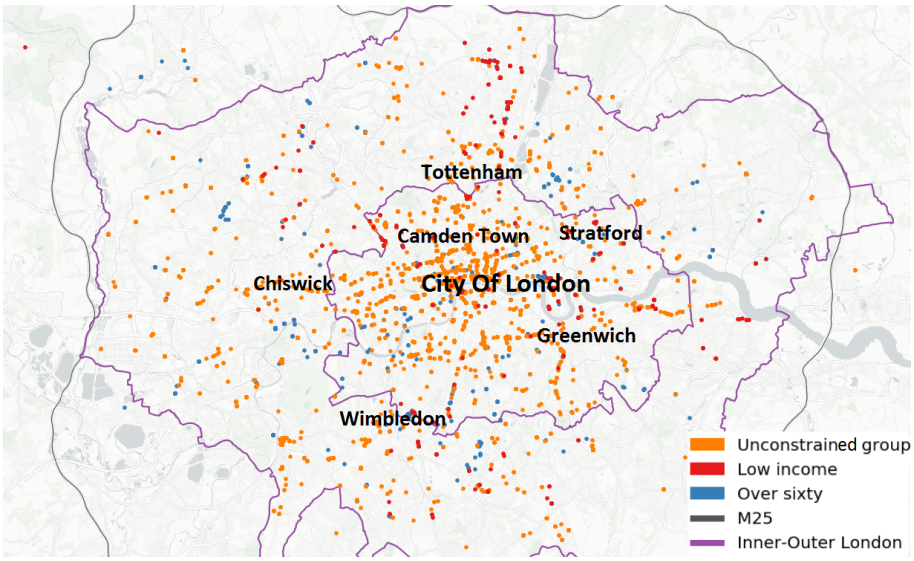


Figure 5.11: Visited places per population group.

$>$ 60 years old population groups is generally concentrated within the boundaries of Inner London, particularly near the City of London . This is not surprising since the majority of employment opportunities are located in this area. On the other hand, the geographical distribution of the low income population group appears to span radially from Inner London, with a significant concentration around the Tottenham area.

5.5.2.3 Ordnance Survey POIs

Similarly to section 5.4, POI data are used for the task of activity type inference. However, contrary to section 5.4 where the nature of the mobility dataset dictated a use of a leisure oriented POI database, this time a more complete POI database was used. This was provided from Ordnance Survey (OS) Points of Interest 2013 dataset. This dataset has a UK wide coverage and consists around 4 million geographic features with location, functional information and addresses, where possible. The database was created in 2002 and is maintained and updated on a continual basis (more than four times a year). The POIs themselves are assimilated from 150 different suppliers and receive regular quality checks on an ongoing basis (Ordnance Survey 2018). This fact makes the OS POI database more complete compared to the Foursquare data used in chapter 6 section 6.3 since it undergoes independent reviews, although the database cannot be considered 100% complete. The POI records, however, have a quality flag attached with them which can be used to inform users about the level of uncertainty associated with their attributes. The complete dataset is grouped into 10 themes: *Acco-*

OS category groupings	**POI categories**	**Example POIs**
Accommodation, eating and drinking	Accommodation; Eating and drinking	Pub, bar, cafe, restaurant etc.
Outdoors and recreation	Gambling; Outdoor pursuits; Sports and entertainment support services; Sports complex; Venues, stage and screen	Stadium; library; theater; park etc.
Education and health	Animal welfare; Education support services; Health practitioners and establishments; Health support services; Primary, secondary and tertiary education; Recreational and vocational education	school; university; hospital; dentist etc.
Retail	Clothing and accessories; Food, drink and multi-item retail; Household, office, leisure and garden; Motoring	supermarket; shop; retail park etc.
Commercial Services	Construction Services; Engineering Services; Consultancies; Personal consuming Services; Repairing	offices; workplaces etc.

Table 5.2: OS POIS used in this case study

modation, eating and drinking, Commercial services, Attractions, Outdoors and Recreation, Education and health, Public infrastructure, Manufacture and production, Ratail, Transport. These are then further disaggregated into more detailed categories that describe specific functions of POIs.

From the 10-fold classification scheme defined by OS (Ordnance Survey 2012), four were considered representative for non-workplace activities (*Accommodation, eating and drinking, Outdoors and recreation, Education and health, Retail*) and one for employment activities (*Commercial services*) (Table 5.2). Note that under this scheme, any employment activities related to education and health will be part of the *Education and health* activity type. Figure 5.12 shows the distribution of the POI categories across central London.

Figure 5.12: OS POI distribution across the study area.

5.5.3 Data preprocessing

This section describes the preprocessing steps followed to derive the observation vector used in the case study of chapter []. These comprised the following tasks:

- Interaction with public transport:
  - Inferring bus and tram alighting stations. This step describes the algorithm for determining the bus stops and tram stations an individual used for alighting.
- Activity type inference:
  - Determination of activity space classification vector. This step describes the data and processes used to derive the vector which is used for activity type inference.

5.5.3.1 Inferring bus boarding and alighting information

As already noted in section 5.5.2, augmentation of Oyster card data with the iBus system resulted in determination of boarding stations for bus journeys. However, this is not the case for alighting stops. As a result, there is a need for a preprocessing step to complete the observed Oyster card/LTDS journeys.

The problem of inferring alighting information from incomplete AFC data is not new and has been studied by many authors using such data in their research. For example, Barry et al. (2009) have built a querying procedure that makes use

of trip-chaining between subsequent ticket validations to infer destinations from boarding only data. In their research, New York City's AFC (called Metrocard) was used. On a similar approach, Zhao et al. (2007) and Wang et al. (2011) have used consecutive trip segments to infer alighting information, the first using data from Chicago's AFC system with the second using Oyster card data. The underlying assumptions for all the above methods are explained in Barry et al. (2002). These are based on the intuition that people tend to use the destination of their previous trip to start the subsequent one. Moreover, people tend to end their last trip of the day at the same station from which the made the first trip. These assumptions are not always correct, as many people could be using a different bus stop to board, rather than the one that they originally exited.

Nevertheless, despite these shortcomings, research has showed that these two simple rules hold for the great majority of users (Gordon 2012, Zhao et al. 2007).

In this research, the trip-chaining approach was followed for determining the exiting station for bus and tram data. The general flowchart of the querying algorithm is shown in Figure 5.13.

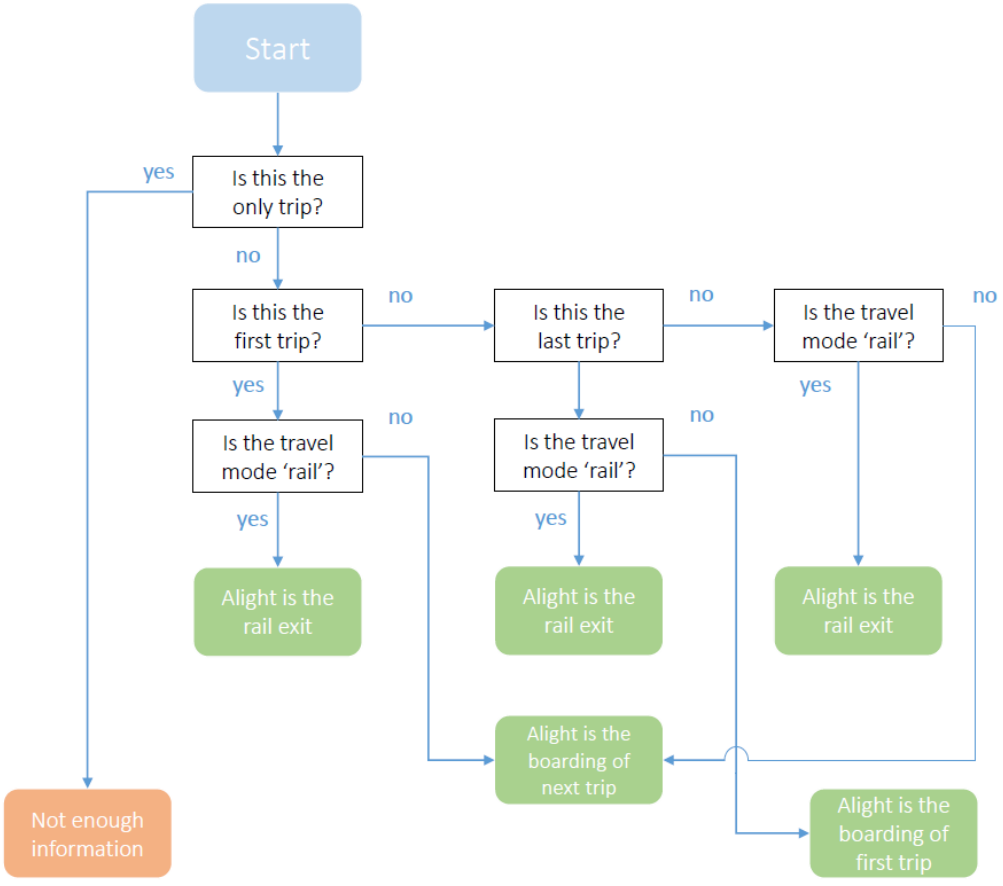


Figure 5.13: Flow chart of destination determination

The algorithm starts by inspecting the public transport trips per day for each individual. Here, a trip is defined by a sequence of Oyster card records that signify the start and end of a particular segment of a journey. If there is only one trip, then the algorithm exits, marking the alighting procedure as unsolvable. If there are more than one trip segments then the algorithm proceeds in checking whether that trip was the first one of the day. If it is, the bus alighting algorithm proceeds in checking the transportation mode in the AFC records. If the transportation mode is rail, then alighting information already exists in the AFC records and the algorithm proceeds to examine the next trip of an individual. If the transportation mode during boarding is bus or tram, the algorithm assigns the boarding point of the next trip as alight, provided that the distance between them is within 8 minutes walking distance along the road network for bus, or 12 minutes along the network for tram (OpenStreetMap was used as the base road network infrastructure) with an average walking speed of 4.5km/h (Evans 2009). If the trip is the last trip of the day for the individual, then the boarding point of first trip of the day is assigned as alight bus stop/rail station, provided the transportation access points satisfies the distance criterion mentioned earlier.

5.5.3.2 Determination of activity space classification vector

The methodology for determining the activity space vector for the task of destination inference is similar to section 5.4.2. This time, however, a series of extra preprocessing steps were undertaken so that features such as location of public transport access points, duration of stay and trip chaining can be taken into account.

Public transport access points catchment area

This catchment area can vary significantly between different individuals depending on factors such as walking speed and distance from a transportation access point to a location of an activity. Within an accessibility framework, these factors are influenced by the personal characteristics of an individual (eg. age, disability, socio-economic status) as well as place based environmental characteristics such as the level of deprivation of an area. For example, in a survey investigating the travel preferences of individuals from social disadvantaged groups (such as lone mothers, people with disabilities and ethnic minority groups) in London, respondents have reported an average walking duration of 15-20 minutes, reaching up to 40 minutes to reach basic activities such as shopping (Wixey et al. 2005). This is a considerable increase compared to the official acceptable estimated walking duration from a transportation access point to a point of interest, which ranges from a maximum

of 8 minutes for bus stops to 12 minutes for rail in London (assuming a fixed walking speed of 4.5 km/hour and not taking into account factors such as walking abilities and environmental factors) (Evans 2009). It is also beyond the limits of minimum acceptable accuracy achievable by using a more detailed activity type categorisation as demonstrated in section 6.3.

The methodology for computing the catchment area that will bound the observation POI vector is similar to that of section 5.4.2. An isochrone network based approach is used with 4 isochrone levels: 5/10/15/20 minutes walking time (assuming 4.5 km/h walking speed). Following the discussion above, in this case study, a different approach was taken to account for the increased uncertainty in activity detection for longer isochrone bands.

Specifically, a linear downweighting scheme was applied to the POI counts in each isochrone band. As such, the POI counts bounded by the 5 minute isochrone remain unchanged, while the 5/10 minute, 10/15 minute and 15/20 minute isochrones are downweighted by 40%, 60% and 80%, respectively. The choice of the three cutoff points was dictated by the need to avoid forming a uniform activity type classification feature space which results from the size of the isochrone area. This way, the contribution of the outer isochrone layers to activities is proportionally reduced with each distance interval.

Differentiating between trip chain and end of journey

Although differentiating between trip chaining and end of journey[3] is straightforward for tube/rail (as this information is readily available in the Oyster card records), this is not the same for journeys made up of bus/tram trip segments. For this task, a minimum duration threshold approach between subsequent trip segments was adopted (Chang & Zhao-Cheng 2016). In particular, a trip is considered part of a journey if it is within the maximum transportation mode interchange times. For rail and tram services, the interchange times were taken from an official request to the service provider submitted in 2015 (Freedom of Information 2015). Figure 5.14 below shows the distribution of interchange times for all rail services within the Greater London Area.

To estimate bus interchange times per bus stop, London's iBus API, was queried for a period of a week (1-7 April 2014) at 10 second intervals and every single response was archived. Part of the API response are predictions of the estimated arrival times for each vehicle and each bus stop across London. Using the time of the API call as reference, the "due to arrive" time was calculated by

[3] A trip is defined as a segment of a journey while a journey is defined as a sequence of trips ending at a destination

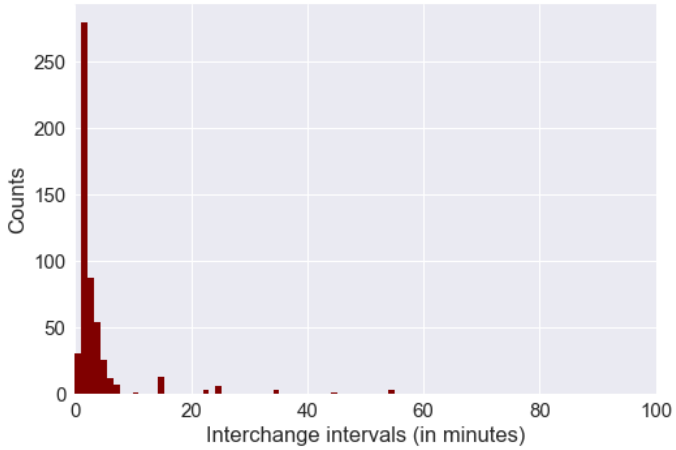


Figure 5.14: Distribution of rail interchange times.

subtracting the reference timestamp from the estimated arrival time. The result was taken as an approximation of the true arrival time of the vehicle at a bus stop. Subtracting subsequent arrival times for every vehicle and every bus stop gives an indication of the distribution of waiting times, which are taken to be a proxy for interchange times. In this way, a more realistic indication of interchange times is obtained, as factors such as delays due to traffic congestion are taken into consideration. Figure 5.15 shows the distribution of bus interchange times for all bus stops in London.

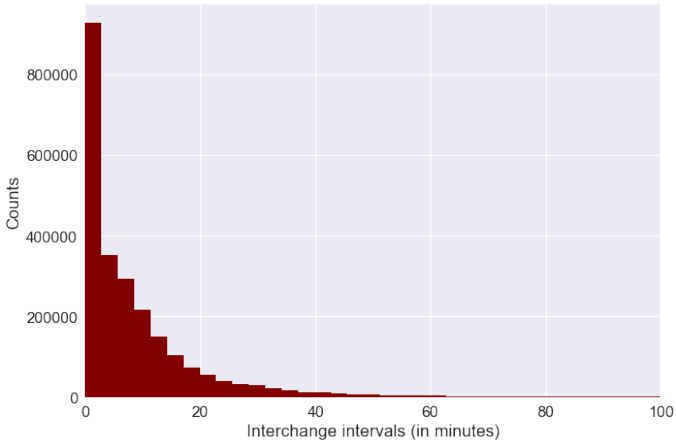


Figure 5.15: Distribution of bus interchange times.

In both cases, the 95th percentile was taken as a cutoff for determining whether an alighting point is considered to be a destination for an activity or an interchange between transportation means. In the case of rail services, this was 15 minutes, while for buses this was 36 minutes. This included roughly 96% of the total Oyster card/LTDS observations. This was used as an end of journey flag in the

Oyster/LTDS dataset used in activity type inference.

Using duration of stay as a weighting function

One of very basic components of an activity (besides location, start time etc.) is its duration. This property has been used to differentiate between activities in the context of rule based approaches (Huang et al. 2010) as well as probabilistic activity clustering approaches (Allahviranloo & Recker 2015, Han & Sohn 2016), particularly for differentiating between employment and non-employment activities.

Having access to the combined Oyster/LTDS data, it is possible to utilise the employment status of individual users to inform the decision threshold with respect to duration. This in turn can provide insights on the nature of employment / non-employment dichotomy.

To do this, individual daily journeys[4] were calculated for each unique Oyster card ID. Then, the duration between individual trips was computed by using only the records that are less likely to belong to an interchange trip. Then the distribution between subsequent daily AFC transactions for individuals with different employment statuses (figure 5.16) was plotted and examined.

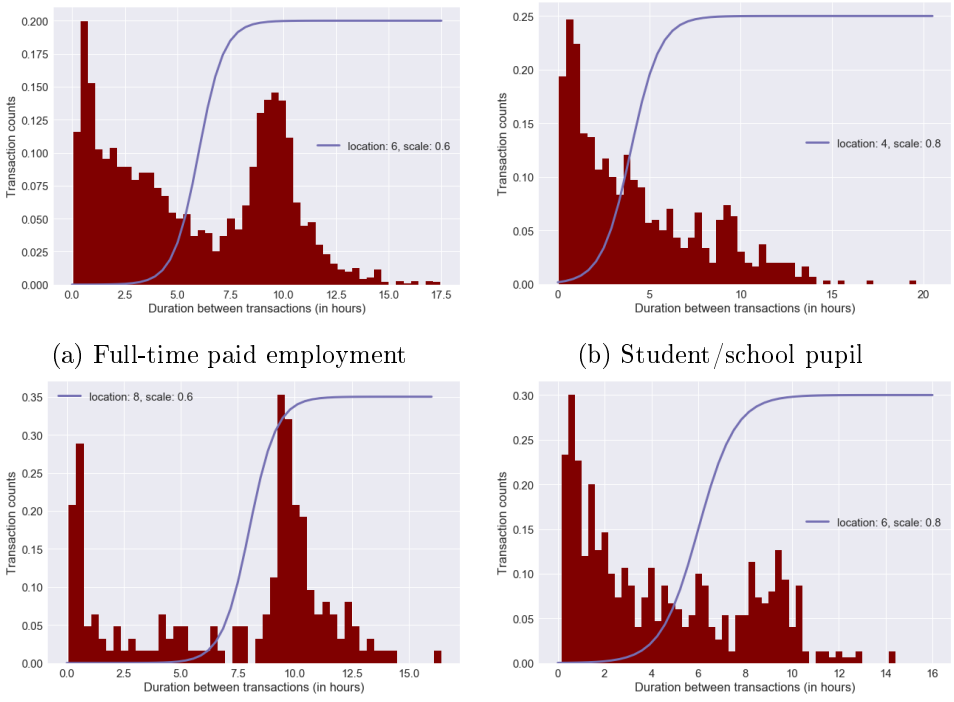
The second mode of the distribution of figure 5.16a, peaking around 9 hours could be attributed to full time employment activities. This is shifted relative to the regular 7.5-8 hour working pattern which could in turn be attributed to the low spatial/temporal resolution of AFC data (not accounting for walking time to and from transportation access points) combined with the way the duration between journeys was calculated (the reference for computing the duration between journeys was the journey's alighting times). The same pattern appears for full-time self-employment (figure 5.16c) with the second mode peaking around 10 hours, reflecting the different working pattern. Individuals that are part-time employed (figure 5.16d) and students (figure 5.16d) display a different working/studying pattern, following less distinct duration cut-off locations which could be attributed to the more flexible nature of part-time employment and education related activity types.

Using the information above, the probability of an activity belonging to "working/studying" category was modulated using the probability density function of a logistic random variable:

[4] A daily journey is defined to be the journey between the first and last "tap" of a day for an individual

$$f(x;\mu,\sigma) = \frac{e^{-\frac{\mu-x}{\sigma}}}{\sigma(1+e^{-\frac{\mu-x}{\sigma}})^2} \tag{5.1}$$

where x is the duration in hours, μ the location parameter and σ the standard deviation (scale). The parameters μ and σ were adjusted to reflect different working assumptions as assessed empirically by the duration of stay distribution of figure 5.16.



(a) Full-time paid employment

(b) Student/school pupil

(c) Full-time self-employment

(d) Part-time paid employment

Figure 5.16: Distribution of duration between transactions for different employment types, with logistic cumulative distribution functions (CDF) overlayed. Note that the histograms were normalised and the CDFs were scaled accordingly.

Equation 5.1 was then used to weight the POI vector corresponding to *Employment/Education* activity types.

5.6 Chapter summary

In this chapter, the datasets used to construct the observation vectors used in the following chapters were introduced. A variety of mobility data were used, having the common characteristic of being (or treated as) unlabelled trajectory points. Initial exploration and preprocessing revealed both systematic and random errors

that diminish the quality of information contained. Attributes such as imbalanced activity type classes, ambiguous transportation mode determination vector, low spatial and temporal resolution and incomplete information, make formulation of a capabilities approach to accessibility model from unlabelled mobility data challenging. As a result, the proposed methodology should be robust enough to provide inferences under the uncertainty introduced by the data idiosyncrasies.

In terms of data preprocessing, the class imbalance characterising the Foursquare dataset was addressed by performing proportional downsampling using UK's labour demand data. The combined Oyster card/LTDS dataset was constructed by performing a one to one record matching for each individual's trajectory between the two datasets. Bus alighting station was imputed using a querying algorithm following the intuition that people tend to use the destination stop as boarding for the following journey. For both Foursquare and Oyster card/LTDS, the POI observation vector was constructed following an isochrone approach using the alighting points as isochrone centroids. For the Oyster card/LTDS in particular, the POIs bounded by the different isochrone bands were proportionally weighted depending on maximum walking time. Finally, the duration of stay was used as a weighting function for differentiating between *Employment/Education* activity types.

Chapter 6

Methodology

6.1 Chapter overview

This chapter develops the necessary components for the formulation of the Capabilities Approach to Accessibility (CAA) model introduced in section 6.4. In the context of the requirements of this thesis described in section 1.1, the modelling approach should allow a) expressing the different accessibility components in a hierarchical way that allows statistical reasoning, b) combining different sources of data (such as passive mobility data and travel survey data) and c) extraction of semantic information (such as activity types and transportation modes) from low level mobility data. To address these requirements, a collection of three interrelated Dynamic Bayesian Networks (DBN) are defined and developed.

In particular, section 6.2 develops a DBN for transportation mode detection that takes into account personal and environmental characteristics. This is accomplished by fusing data from travel surveys and machine generated mobility data in a complementary way. The results of this section have been published in the article "Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics" in the journal Transportation Research Part C: Emerging Technologies (Bantis and Haworth 2017).

Section 6.3 develops a DBN for inferring activity types accounting for the potential activities an individual might be performing at a destination. Given the low spatiotemporal resolution of data generated by service providers (e.g. transportation or mobile network operators), this section performs a robust assessment of the degree of accuracy achievable. The findings of this section have been published in the article "Non-Employment Activity Type Imputation from Points of Interest and Mobility Data at an Individual Level: How Accurate Can We Get?" of the ISPRS International Journal of Geo-Information (Bantis and Haworth 2019).

Finally, section 6.4 consolidates the modules developed in sections 6.2 and 6.3

and defines a DBN structured around the Capabilities Approach (CA). Using this model, the link between social exclusion and transport disadvantage in investigated in the case study of chapter 7. Parts of this chapter have been published in the article "Assessing transport related social exclusion using a Capabilities Approach to accessibility framework: A dynamic Bayesian network approach" of journal Journal of Transport Geography (Bantis and Haworth 2020).

Figure 6.1 shows the structure of the framework. The models of sections 6.2 and 6.3 are responsible for fusing, pre-processing heterogeneous data sources and extracting semantic information from passive mobility data, while the model of section 6.4 combines the outputs of the two modules and defines the accessibility model structured around the CA.

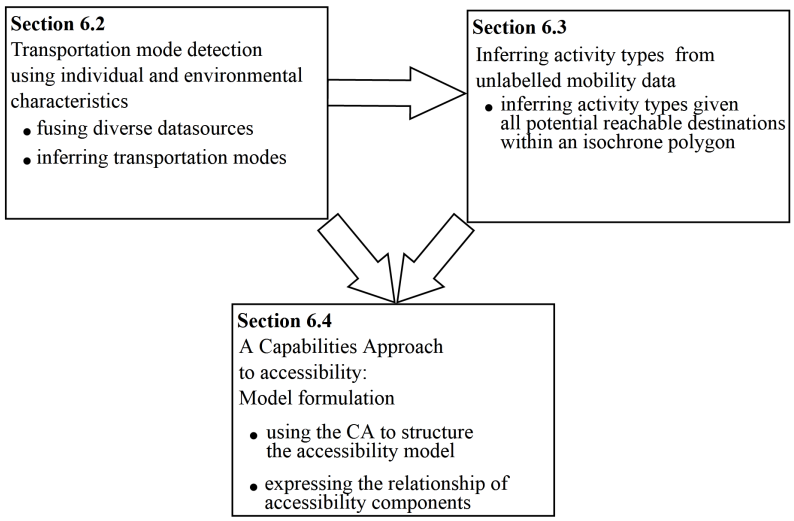


Figure 6.1: Roadmap for developing the Capabilities Approach to Accessibility (CAA) model.

6.2 Transportation mode detection using individual and environmental characteristics

This section describes the DBN used for combining diverse datasources and inferring transportation modes from unlabelled mobility data. In particular, section 6.2.1 describes how the personal and environmental characteristics are included in the DBN, as well as the way the dynamic element is expressed in the model specification. Furthermore, it describes the data augmentation strategy for fusing travel survey and passive mobility data in the modelling process. The performance and feasibility of this specification is demonstrated within the context of

using mobility data from a custom developed smartphone app (see section 5.3) to simultaneously infer individual transportation modes used and assess the effect of personal and environmental characteristics in the mode choice (section 6.2.2).

6.2.1 Model specification

The model is conceptually represented by the the graph of figure 6.2. Within the context of transportation mode detection using unlabelled mobility data, the potential modes used by an individual is determined using speed readings (see section 5.3).

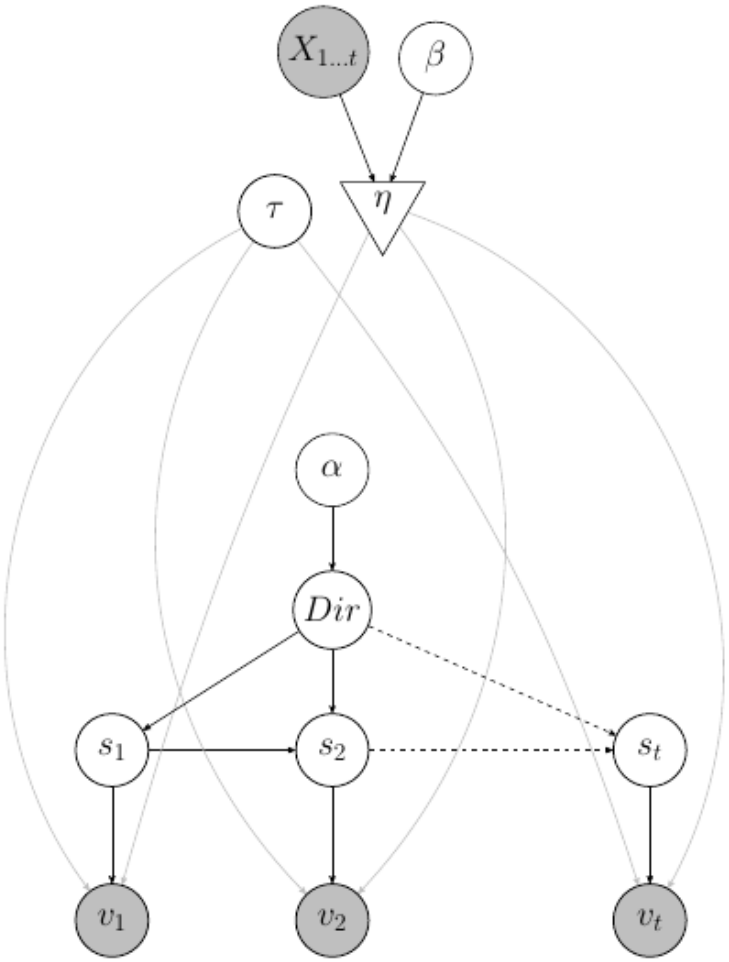


Figure 6.2: Schematic representation of the model

Table 6.1: Description of nodes in Figure 6.2

Node	**Description**
$v_{1...t}$	Speed
$s_{1...t}$	Transportation mode states
Dir	Dirichlet distribution
α	Concentration parameter vector for the personal preferences
τ	Precision vector of speed node
η	Deterministic function of mean for the speed node
β	Coefficient of external covariates
$X_{1...t}$	External covariates

The different transportation modes were modelled using a categorical probability distribution, having outcomes as described by a predetermined set of transportation modes such that $\sum s^{\kappa} = 1$ where s^{κ} are the event probabilities. For this research four different categories were used: being stationary, walking, riding the bus/driving a car and travelling by rail. The emission probabilities $P(v_t \mid s_t)$ were modelled as as a mixture of Gaussian distributions representing the range of velocities each travel mode can take (Patterson et al. 2003, Liao, Patterson, Fox & Kautz 2007).

A common problem encountered with the above approach during inference is related to the identifiability, or label-switching, between the candidate classes. This refers to permuting the subscripts of the mixture components without changing the likelihood in such as a way that the interpretability of inferred classes is lost (Congdon 2010). Various strategies have been suggested in the literature to deal with this problem, from imposing an sorting structure (ascending or descenting) on the Gaussian components (Zucchini & MacDonald 2009), to the use of informative priors (Congdon 2010). Due to its simplicity, a sorting structure has been applied in this study such that $\mu_t^1 < \mu_t^2 < ... < \mu_t^{\kappa}$ for $\kappa \in \{stationary,\ walk,\ bus,\ rail\}$. The initial probability of using a particular transportation mode $P(s_0)$ was evaluated following the condition for a stationary Markov Chain following (Zucchini & MacDonald 2009). This states that the vector $\boldsymbol{x}$ is the stationary distribution for the stochastic matrix $\boldsymbol{P}$ if and only if:

$$\boldsymbol{x}(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{U}) = \boldsymbol{1} \tag{6.1}$$

where $\boldsymbol{x}$ is the vector of non negative elements of the stationary distribution, $\boldsymbol{I}$ is the $\kappa \times \kappa$ identity matrix, $\boldsymbol{P}$ is the transition matrix and $\boldsymbol{U}$ is an $\kappa \times \kappa$ matrix with all elements equal to one.

Two assumptions are made at this point which are important for constructing the CAA model presented in the following sections:

1. Socio demographic characteristics have a persistent effect on the ability of individuals to transition from one transportation mode to another

2. Environmental characteristics have a non persistent effect on the ability of individuals to transition from one transportation mode to another

The first assumption is related to the personal circumstances of an individual when switching between different transportation modes. For example, an individual with disabilities might prefer to use a transportation mode that is more accessible compared with the other, and this preference is assumed to be constant regardless the data one is observing.

On the other hand, external factors, such as whether an individual is located within the catchment area of a bus or a rail station, are assumed to change throughout an individual's trajectory. For example, an individual moving within the radius of bus stops, is more likely to be using a bus.

6.2.1.1 Including external covariates

For this study, a 30 meter radius around bus stops and rail stations was taken as threshold to define the binary covariates depending on whether a person is located within, or outside this radius.

This threshold corresponds to a compromise between the maximum achievable accuracy when the API is using a WiFi/Cell tower level accuracy and the minimum achievable accuracy when using the GPS sensor. Other spatially varying covariates that are assumed to influence an individual's mobility can be included. These could range from socio economic features such as crime levels, to features that characterise the aesthetic quality of a route (Evans 2009). For this study, the Index of Multiple Deprivation (IMD) was taken as a proxy for the level of attractiveness of an area. IMD is an index made up of seven sub-indices relating to features such as income level, employment, health, education skills, barriers to housing and services, crime and living environment. The index ranks the different UK census areas from most deprived to least deprived (UK Government 2015).

For this study, the values were normalised to have zero mean and unit variance to assist inference as the scale difference between IMD and proximity covariates

ranges from one to two orders of magnitude. Figure 6.3 below shows the location traces of the mobility scooter participant together with the levels of IMD for each census area.

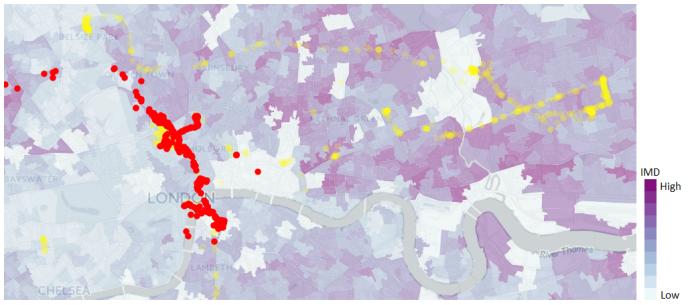


Figure 6.3: IMD overlaid with a participant's traces. The value of the covariate changes according to the census area he/she is located. The red traces correspond to the wheelchair participant, while the yellow to the crutches participant.

6.2.1.2 Including personal characteristics

Personal characteristics depending on age and disability were used to shape the prior belief of a person using one transportation mode over another. The choice of the shape of prior distribution that can be used to reflect the prior belief has received much attention in literature. Three approaches to specifying prior distributions can be found (Gelman et al. 2013): Uninformative, informative and weakly informative prior distributions. Uninformative prior distributions are constructed in a way that have minimal impact on the posterior quantities, so that inferences are dominated by information related to the observed data. A related concept is weakly informative priors with the difference that in this case, the prior distribution contains enough information to keep inferences within reasonable bounding values without capturing any explicit knowledge about the state of the model. Informative prior distributions on the other hand, are constructed to reflect the state of knowledge about the possible values of the model parameters before observing any data. For this case study, an informative approach was followed, where the prior belief was expressed by drawing samples from an asymmetric Dirichlet prior distribution during inference.

The choice of a Dirichlet distribution prior is a natural choice for this problem given the fact that it is the conjugate prior of the categorical distribution of transportation states. This section describes the approach for determining the concentration parameters of the Dirichlet distribution.

The vector of values of the concentration parameters was used to control the

Table 6.2: p-values of chi-squared test between transportation modes and sociodemographic variables using the LTDS dataset.

	Bus	**Walk**	**Rail**	**Disability**	**Age**	**Income**
Bus	NA					
Walk	2.13e-219	NA				
Rail	2.55e-242	2.13e-219	NA			
Disability	3.14e-13	0	6.39e-210	NA		
Age	2.21e-52	4.70e-117	5.049e-312	2.93e-284	NA	
Income	9.73e-121	0.1080	6.16e-158	6.63e-125	4.07e-167	NA
Sex	9.64e-11	0.1962	1.06e-29	0.0034	0.2626	9.97e-15

level of prior belief in the preferences of an individual towards the transportation modes. Smaller ($0 < \alpha < 1$) values of α express less uncertainty in the preference of a transportation mode over the other. On the other hand, bigger values ($\alpha > 1$) express more uncertainty on the preference of an individual for a transportation mode. A concentration parameter vector with unit values would represent complete ignorance over the preferences of a user, or a user with no particular preferences. Most commonly, values between 1-5 are used if the concentration parameters are assumed to be pre-set, or they can be assigned a prior distribution, most commonly a Gamma distribution (Congdon 2003).

In this study, the calculation of different concentration parameter priors was based on the London Travel Demand Survey (LTDS) dataset described in section 5.5.2.

Examining the pairwise differences between the frequencies of journeys using different transportation mode and variables such as age, income, sex and disability using a Pearson's chi-squared test for the LTDS data (Table 6.2), one could see that for most variables there is a significant difference between the frequency of transportation mode use and socio-demographic characteristics. This suggests an overall strength of association between these variables. Exceptions are the variables income and sex in relation with walking.

The overall workflow of concentration parameter calculation is shown schematically in Figure 6.4.

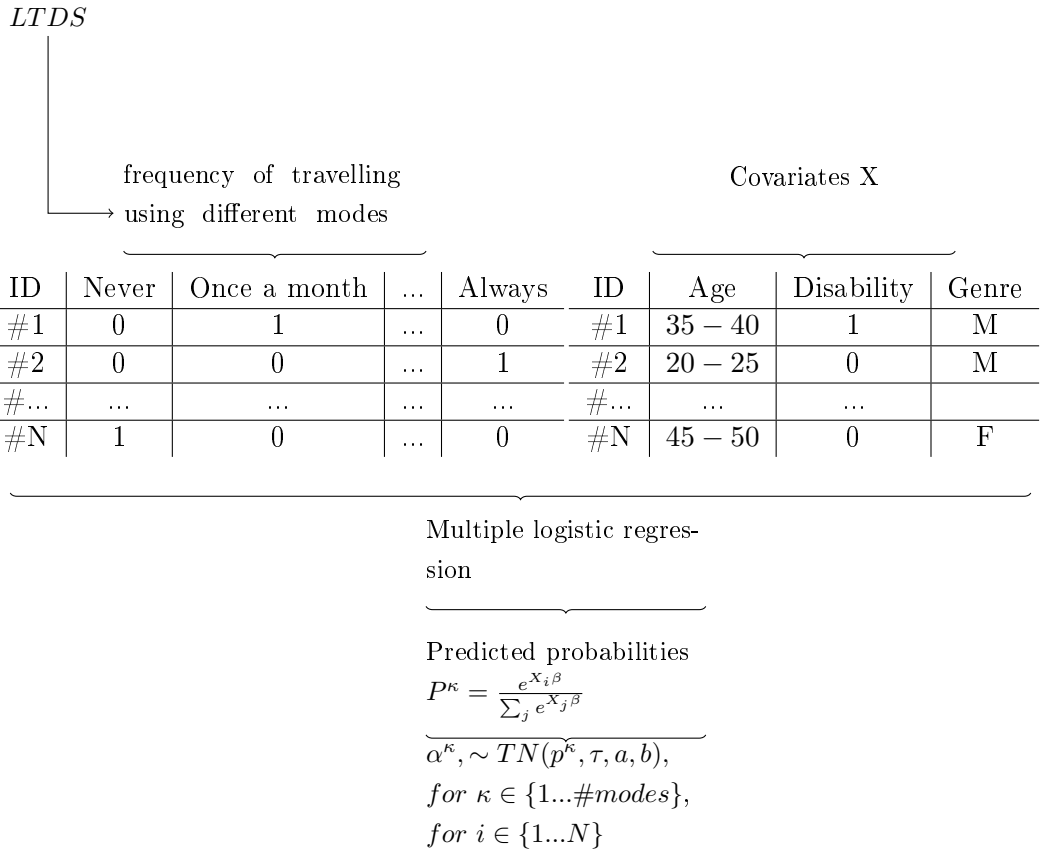
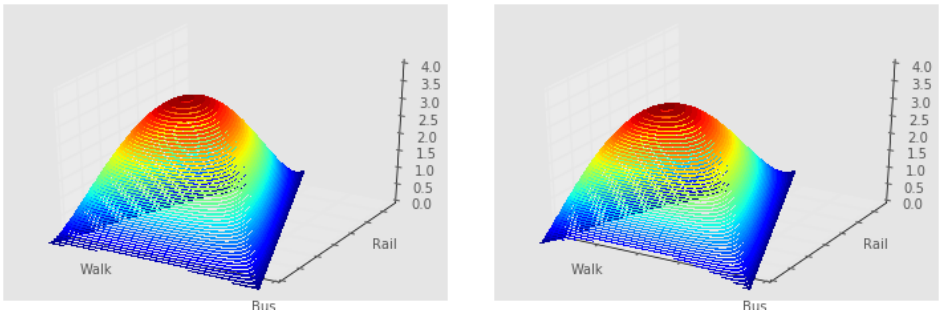


Figure 6.4: Concentration parameter calculation work-flow using LTDS data

The participant responses from LTDS datasets to walking, using the bus and rail transportation modes were dummy coded into multiple binary variables based on the frequency of use. The breakpoint condition in the coding procedure was the use of the respective transportation mode for more than once per month. The resulting data were then used in a multiple logistic regression model with independent variables being age and binary coded disability status and sex of the individuals. The predicted probabilities of using each transportation mode were then calculated using the actual age and disability status for each of the two participants in this study. The resulting values were used when modelling the mean parameter in the truncated normal distributions before including them as concentration parameters in the calculation of the Dirichlet prior. This was to allow for increased uncertainty between different transportation modes while ensuring that the values drawn were all positive. The stationary state was given a value of 1 for all participants reflecting lack of knowledge for this specific state.

The benefits of the above procedure are three-fold. First, by injecting prior knowledge in the model, the inference procedure becomes more robust as the posterior is weighted away from unlikely values as determined by past studies.



(a) Male aged between 20-39, disabled, $\alpha = [1.93, 1.75, 1.64]$

(b) Female aged between 40-59, disabled, $\alpha = [1.91, 1.75, 1.44]$

Figure 6.5: Dirichlet distribution results

Second, this procedure allows the determination of the extent of influence of socio-demographic characteristics shared amongst population groups when assessed at the individual level. Third, in this way, a framework for combining information from different sources is introduced, allowing for a more detailed representation of mobility behaviour.

The Figure 6.5 below shows the resulting Dirichlet distributions for the two participants in the study. In this figure, each corner of the triangle corresponds to a potential transportation mode, while the z axis corresponds to the Dirichlet probability mass.

More formally, the final model is defined in equation 6.2:

$$
\begin{aligned}
P(s_t^\kappa | s_{t-1}^\kappa) &\sim Cat(\boldsymbol{p}), \\
P(m^\kappa) &\sim N(0, 10^{-3}), \\
\eta_t &= \beta_0 + \boldsymbol{\beta'} \boldsymbol{X}, \\
P(\mu^\kappa) &= \eta_t m^\kappa, \\
P(v_t^\kappa | s_t^\kappa) &\sim N(\mu^\kappa, \tau^\kappa) \; where \; \mu^1 < \mu^2 < \mu^3 < \mu^4
\end{aligned}
\tag{6.2}
$$

The hyperpriors in this model were:

$$
\begin{aligned}
P(\boldsymbol{p}) &\sim Dir(\boldsymbol{\alpha}), \\
P(\alpha^\kappa) &\sim TrN(a^\kappa, 0.01, 0.1 < bound < +\infty), \\
P(\beta_{1..\#covariates}) &\sim N(0, 10), \\
P(\tau^\kappa) &\sim Gamma(0.001, 0.001)
\end{aligned}
\tag{6.3}
$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and β_0 is the intercept term both

are assumed to be distributed as a normal distribution with mean 0 and precision 10^{-5}, and $\mathbf{X}$ is the matrix of external covariates.

6.2.2 Results

This section provides the results of travel mode detection using the specification described in the previous sections. All inferences were carried out within a Bayesian framework, using Markov Chain Monte Carlo (MCMC) methods. A MCMC is a numerical approximation algorithm that attempts to approximate the posterior distribution by drawing sequential samples, with the distribution of the sample draws depending on the last sample drawn. The sequential nature of the samples, allows the approximate distribution to improve at each step of the simulation and converge towards the "true" posterior distribution (Gelman et al. 2013). The fundamental desired property of an MCMC algorithm is to create a Markov process whose stationary distribution is a specified posterior, so that long enough simulations generate samples which are close enough to the target distribution. Many different MCMC algorithms have been designed with this property in mind, with Metropolis-Hastings being one of the most widely used ones. The algorithm uses a proposal distribution $g(x^*|x)$ to update the state of the variable x by drawing candidates x^* from $g(x^*|x)$ and computing the acceptance probability $\alpha(x^*, x) = min(1, \frac{g(x|x^*)\pi(x^*)}{g(x^*|x)\pi(x)})$ where $\pi(x)$ is the target distribution (Neal 2000). For this case study, different MCMC sampling schemes were employed for the different stochastic nodes of the Bayesian network. Specifically, for the categorical nodes ($P(s_t^\kappa \mid s_{t-1}^\kappa)$) a discrete Metropolis sampling scheme, and for the continuous nodes ($P(m^\kappa), P(\mu^\kappa)$, $P(\boldsymbol{p})$, $P(\alpha^\kappa)$, $P(\beta_{1..\#covariates})$, $P(\tau^\kappa)$) a combination of Metropolis-Hastings, Adaptive Metropolis and Gibbs (Hit and Run sampler) (Brooks et al. 2011).

For all models, 5×10^5 iterations were used to approximate the unknown parameters. Convergence was assessed using visual methods and Geweke's convergence diagnostic. The first involves inspecting the MCMC chains for non-stationarity while the second compares the mean and variance using a Z-score test between the first and last segment of the Markov chain to assess whether there are statistically significant differences (Geweke et al. 1991) (figure 6.6):

$$z = \frac{\bar{\theta}_a - \bar{\theta}_b}{\sqrt{Var(\theta_a) + Var(\theta_b)}} \tag{6.4}$$

where α and β are the first and last part of the chain respectively. For this application, this was taken to be 10 and 50% respectively.

The tests indicate that convergence has been achieved although additional

samples would have improved the characterisation of posterior quantities. This is particularly evident for the crutches user where the mixing of the samples was slower compared to the wheelchair user. Increasing the standard deviation for the latter to allow for more values to be rejected could improve mixing in the latter case. The acceptance ratio for all parameters for both users was found to be satisfactory, within the range of 0.2-0.245. In figure 6.7 below, the convergence diagnostics are shown for the v node of figure 6.2.

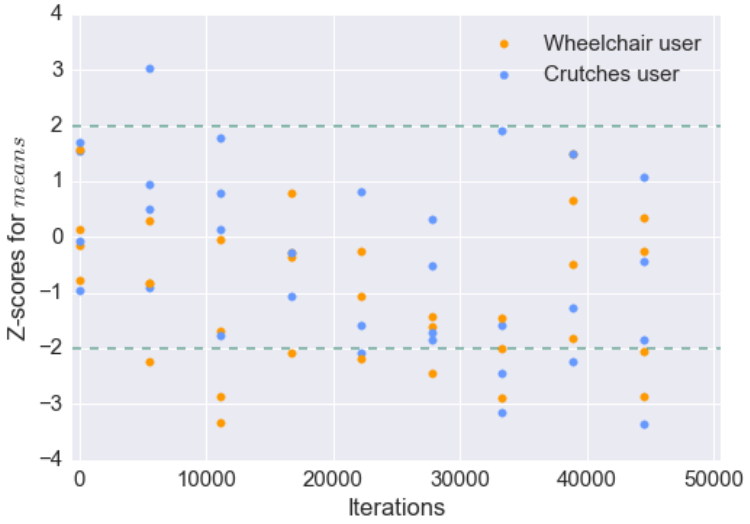


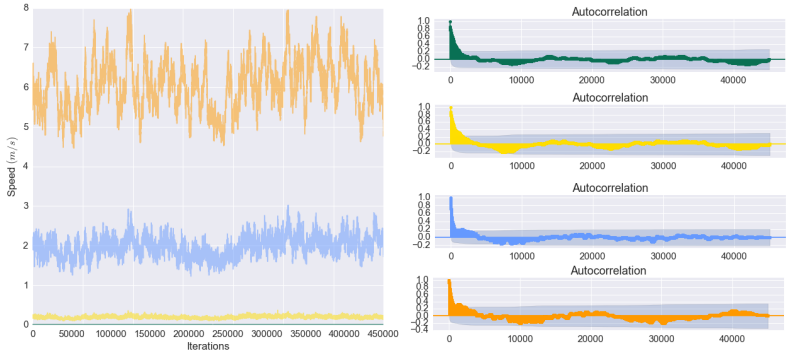
Figure 6.6: Geweke Z-scores for the speed nodes. The majority of the samples are within two standard deviations from the mean of the first and the last segment of the MCMC chain.

The posterior distributions of the inferred speeds for the two participants are shown in figure 6.8 below. As it can be seen, the inferred speed is considerably different, especially for the walking mode. This is to be expected considering the fact that the two participants use different mobility aids when travelling without using car/public transport.

The dynamics of the participants interactions with the different transportation modes was captured in the transition matrices. These are stochastic matrices with each row representing a categorical distribution of switching between modes. This corresponds to the transition probability $P(s_t^\kappa \mid s_{t-1}^\kappa)$ of the model. Figure 6.9 below shows the posterior quantities of the transitions probabilities between different transportation modes.

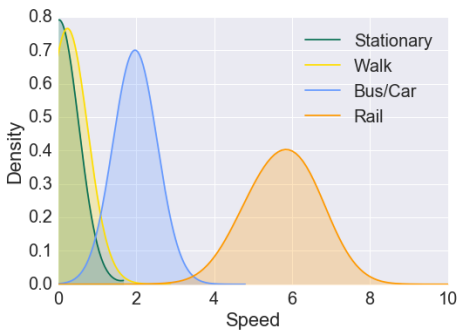
For both participants, the effect of external factors on their movements was found to be either very small or statistically non-significant, given that the zero value is contained within the 95% credible intervals, and this was true for different β covariates. The exception was the wheelchair user, where the IMD had a positive

(a) MCMC traces and autocorrelation plots for the wheelchair user

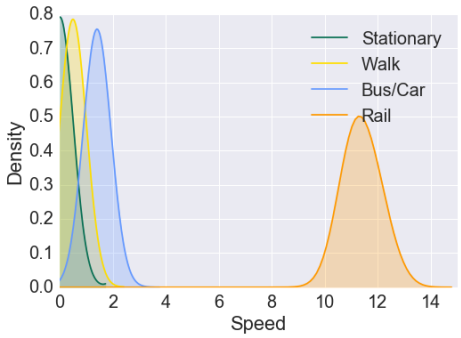


(b) MCMC traces and autocorrelation plots for the crutches user

Figure 6.7: MCMC traces and autocorrelation plots for the two participants. The slow mixing of the crutches user can be seen from the tendency of the MCMC chain to make small jumps when proposing new speed values.

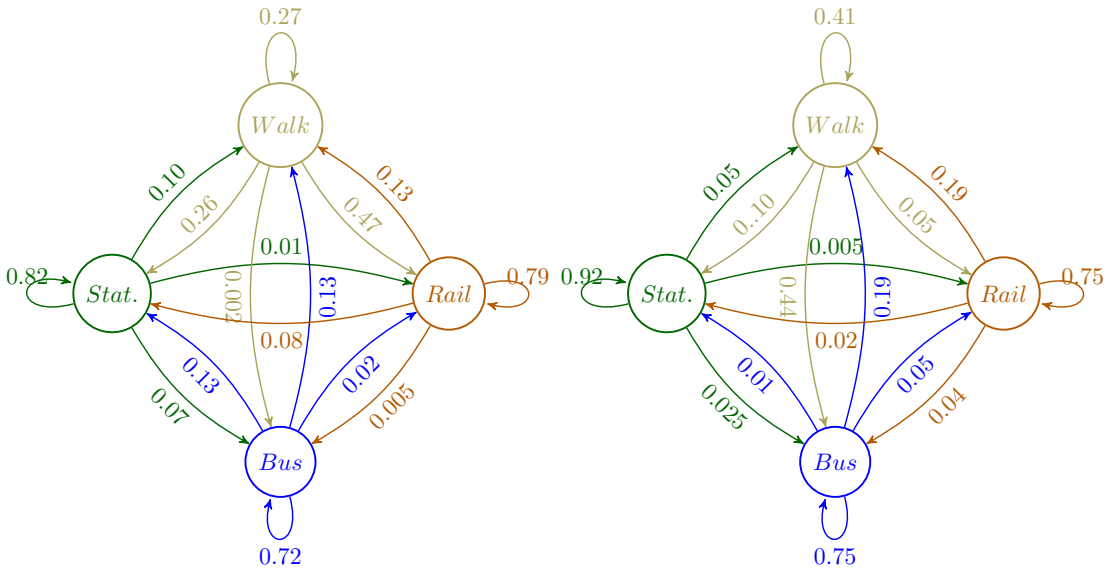


(a) Posterior speeds for the crutches user



(b) Posterior speeds for the wheelchair user

Figure 6.8: Posterior quantities of the speed node (v). The histograms are normalised. The unit of measurement is m/s.



(a) Posterior mean transition probabilities for the crutches user

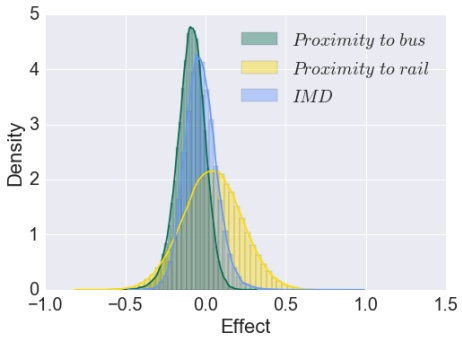
(b) Posterior mean transition probabilities for the wheelchair user

Figure 6.9: Transition probabilities for the two participants. The color coding corresponds to each row of the transition matrix with values as indicated from the corresponding arrows.

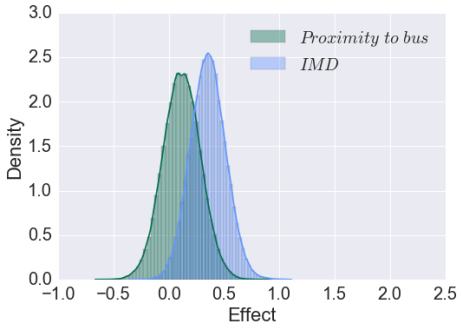
effect of 0.35 (Figure 6.10). These correspond to the β nodes of the model in figure 6.2.

Looking at the internal effects as expressed by the concentration parameters of the Dirichlet distribution, (the α node of figure 6.2), one observes that the values have shrunken towards values less than one, concentrating the Dirichlet distribution towards each individual categorical node and reflecting increased certainty of preferences of one transportation mode over the other (figure 6.11).

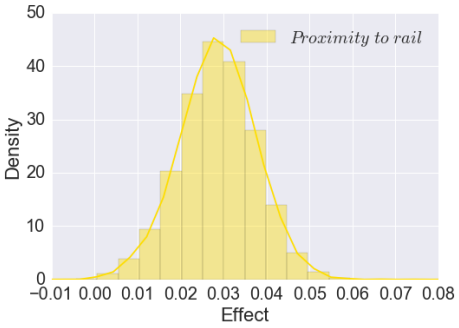
For the s node in figure 6.2 model the posterior classification travel mode detection accuracy was assessed using the participant's own travel mode labelling. Figure 6.12 below shows the posterior median quantities for all data points, categorised by day.



(a) External factors effect for the crutches user

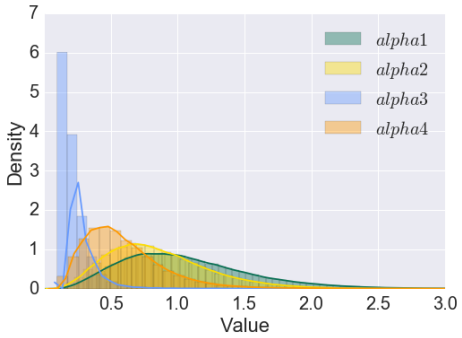


(b) External factors effect for the wheelchair user

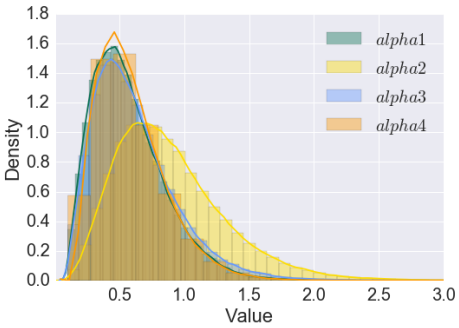


(c) External factors effect for the wheelchair user

Figure 6.10: Posterior quantities for the IMD, proximity to bus stops and proximity to rail stations.

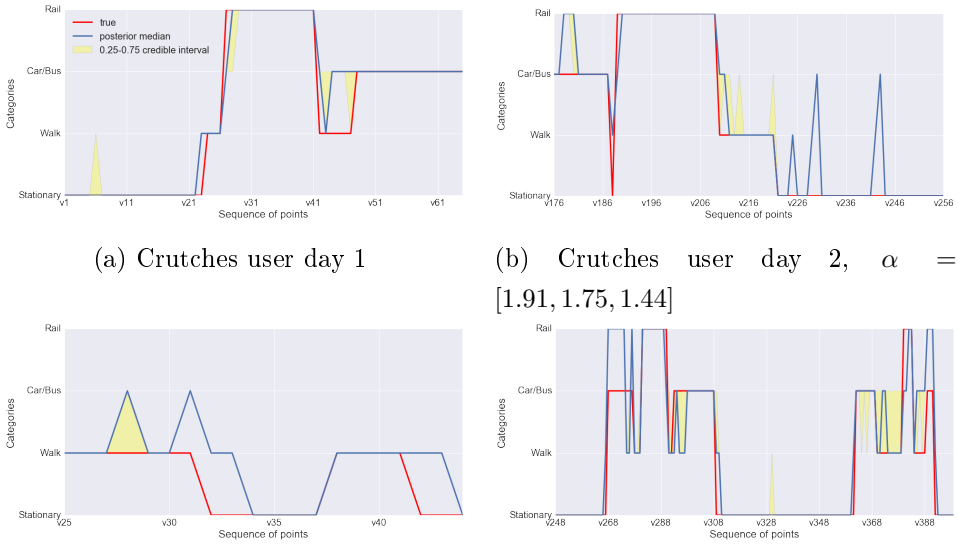


(a) Concentration parameters for the crutches user



(b) Concentration parameters for the wheelchair user

Figure 6.11: Posterior quantities for the truncated normal priors of the Dirichlet concentration parameters. The color coding represents the corresponding categorical distributions and is the same as Figure 6.9.



(a) Crutches user day 1

(b) Crutches user day 2, $\alpha = [1.91, 1.75, 1.44]$

(c) Wheelchair user day 1

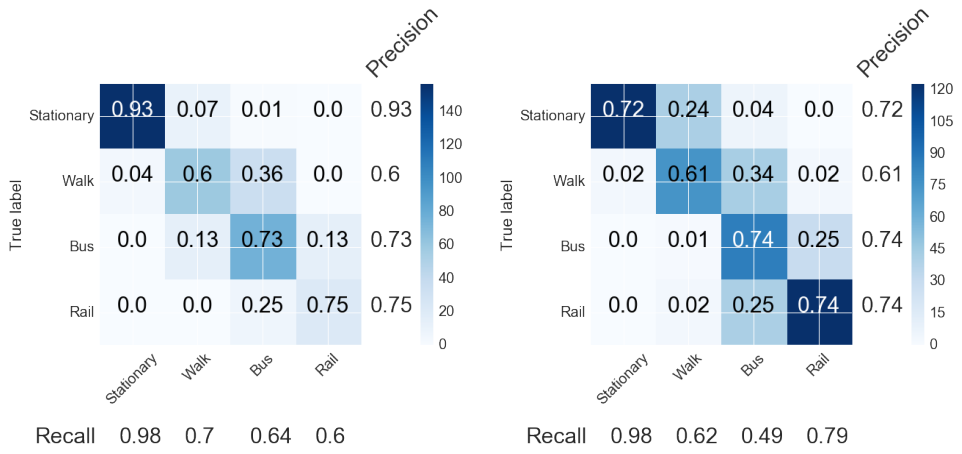
(d) Wheelchair user day 2

Figure 6.12: Posterior medians of the categorical node (s) of the model. The red line correspond to the self-labeled value while the blue line corresponds to the inferred quantities. The yellow faces are the 95% credible intervals of the MCMC simulation.

6.2.2.1 Performance evaluation

The performance of the proposed method was compared to other popular classification algorithms, namely Random Forests (RF), Support Vector Machines (SVM's) and Multilayer (MPL). It is important to notice that, contrary to the aforementioned classifiers, the proposed method is essentially an unsupervised classification procedure, and as such a training step is not needed. For this task, 70% of the data-sets were used during the training procedure, and 30% for testing. The benchmark for comparison was the participants self-labeled true states.

The overall classification accuracy using the proposed method is illustrated in the confusion matrix (figure 6.13). It can be seen that missclassification mostly occurred between the walk and the bus travel modes. This can be explained by considering the lower accuracy of data generated by mobile phone API's, together with the low mean speed of travel for buses in peak hours in London, which could be as low as 6km/h (Transport for London | Every Journey Matters 2017). The same holds true for missclassification artefacts between walking and stationary states, which could be attributed to the effects of "sudden jumps" and "drift" in location. The overall accuracy was 71% for the wheelchair user and 78% for the crutches user.

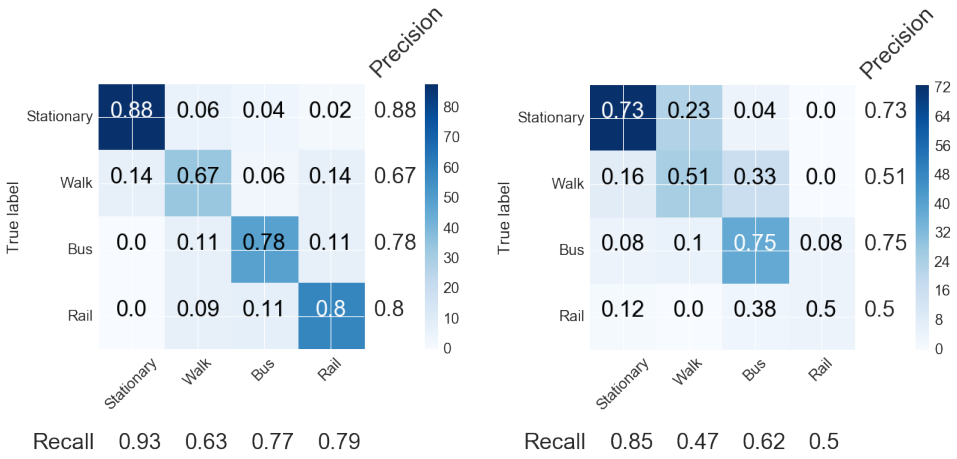


(a) Confusion matrix for the crutches user

(b) Confusion matrix for the wheelchair user

Figure 6.13: Confusion matrices between the self labeled data and the inferred transportation modes for the proposed method. The color-bar corresponds to the number of data points.

Next, a RF classifier was employed. Maximum accuracy was achieved for training 10 trees in the forest. No restrictions were placed for the maximum number of features for each individual tree. Looking at the results of the RF classifier, missclassification is more profound for the walking mode. This is especially true for the wheelchair user dataset which proved to be challenging in terms of classification performance. The overall accuracy was 79% for the crutches user and 67% for the wheelchair user.



(a) Confusion matrix for the crutches user

(b) Confusion matrix for the wheelchair user

Figure 6.14: Confusion matrices between the self labeled data and the inferred transportation modes for the RF classifier.

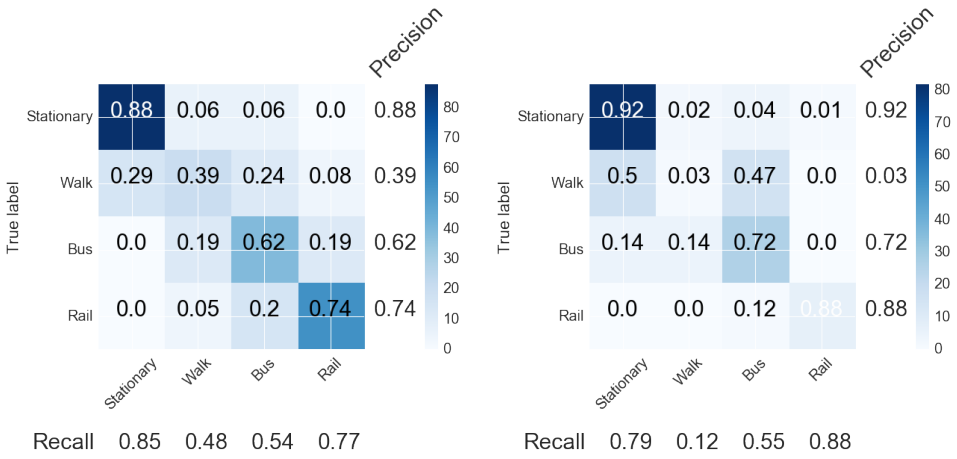
Next, a SVM classifier was employed with an exponential kernel. Maximum accuracy was achieved with a penalty parameter of 1 and a γ value of $\frac{1}{\#features}$ The algorithm was found to perform comparatively well to both RF and the proposed method. The accuracy for the wheelchair user was 62% while for the crutches user was 71%.

(a) Confusion matrix for the crutches user

(b) Confusion matrix for the wheelchair user

Figure 6.15: Confusion matrices between the self labeled data and the inferred transportation modes for the SVM classifier.

Finally, a Multilayer Perceptron (ANN) was employed using backpropagation for the training procedure. The total number of hidden layers that yielded maximum accuracy was 15. The algorithm performed comparably to both the proposed method and RF and overperformed the SVM. For this classifier the accuracy for the wheelchair user was 69% while for the crutches user was 70%.



(a) Confusion matrix for the crutches user

(b) Confusion matrix for the wheelchair user

Figure 6.16: Confusion matrices between the self labeled data and the inferred transportation modes for the MLP classifier.

To assess whether the classification results between the different methods were statistically significant, a chi-squared test was carried out between the classification results of the proposed method and the results of RF, SVM and MLP classifiers. The null hypothesis is that the difference in the classification results could be generated by chance alone. Looking at the p-values, the proposed method produced statistically significant results for all of the transportation modes for the wheelchair user and for nearly half the transportation modes for the crutches user. The corresponding chi-squared statistics and p-values are presented in tables 6.3 and 6.4 below:

Mode	**Statistic**	**RF**	**SVM**	**MLP**
Stationary	chi-sq	7.4976	2.9728	7.2529
	p-value	0.0576	0.2261	0.0266
Walk	chi-sq	29.6063	29.6905	28.929
	p-value	1.67E-06	1.60E-06	2.32E-06
Bus	chi-sq	0.4605	1.9966	2.28382
	p-value	0.7943	0.57310	0.31920
Rail	chi-sq	5.48766	5.48766	1.72545
	p-value	0.06432	0.0643	0.4220

Table 6.3: Chi-squared statistic and p-values for the crutches user.

Mode	Statistic	RF	SVM	MLP
Stationary	chi-sq	10.2118	6.7393	21.7692
	p-value	0.0060	0.0344	7.29E-05
Walk	chi-sq	30.2307	30.4187	67.8636
	p-value	1.23E-06	1.13E-06	1.22E-14
Bus	chi-sq	13.0388	71.4039	36.9357
	p-value	0.0045	2.14E-15	4.75E-08
Rail	chi-sq	12.2690	21.1491	0.8027
	p-value	0.0021	9.80E-05	0.6694

Table 6.4: Chi-squared statistic and p-values for the wheelchair user.

Finally, to assess the generalisation of the model to different datasets, the proposed method was employed to a GPS dataset of 5 individuals. The temporal resolution of this dataset was 60 seconds of each subsequent GPS point, while the spatial accuracy according to the horizontal dilution of precision value was ≤ 2.5 which translates to a good accuracy level for most applications (Langley et al. 1999). The temporal domain of this dataset spanned for over a week for all participants. Since information about the individual socio-demographic characteristics of participants was unavailable for this sample, an uninformative Dirichlet prior distribution was used for modelling the effect of personal preferences. The performance of classifiers was tested against the individuals self-labelled data. The hyper parameters of SVM, RF and MLP were tuned for different values using the exhaustive grid search method. The results are presented in the table below along with the corresponding results of SVM, RF and MLP classifiers. As it can be seen, the proposed method performed comparably when compared to SVM, RF and MLP classifiers.

Table 6.5: Performance evaluation using a GPS sample of 5 individuals.

Classifier	**Stationary**	**Walk**	**Bus/Car**	**Rail**	*Recall*
	0.94	0.05	0.0	0.0	0.94
Pr. method	0.28	0.64	0.03	0.05	0.65
	0.03	0.04	0.92	0.01	0.92
	0.02	0.04	0.01	0.94	0.94
Precision	0.94	0.65	0.88	0.91	*Acc: 0.90*
	0.91	0.05	0.03	0.01	0.91
RF	0.36	0.49	0.12	0.02	0.49
	0.18	0.08	0.65	0.08	0.65
	0.07	0.05	0.16	0.72	0.72
Precision	0.92	0.53	0.44	0.84	*Acc: 0.82*
	0.89	0.09	0.02	0.01	0.89
SVM	0.29	0.61	0.03	0.08	0.61
	0.04	0.08	0.77	0.11	0.77
	0.09	0.02	0.09	0.81	0.81
Precision	0.95	0.43	0.53	0.79	*Acc: 0.85*
	0.91	0.05	0.03	0.01	0.91
MLP	0.36	0.49	0.12	0.02	0.49
	0.18	0.08	0.65	0.08	0.65
	0.07	0.05	0.16	0.72	0.72
Precision	0.92	0.53	0.44	0.84	*Acc: 0.82*

6.2.3 Discussion

The posterior quantities for the speed of the two participants in the study were found to be different when compared with each other. This can be attributed to numerous reasons. As already mentioned, the mobility aids each participant is using are different, influencing the walking speed in different ways. In addition, each participant is using different rail transportation modes, namely the London Overground rail service for the crutches user and the National rail services for the wheelchair user. These two services have very distinct speed signatures, the first one being a city wide transportation mode with more frequent stops while the latter being used for intercity travels with fewer intermediate stops. Looking at the state transition probabilities, a detailed insight on the way the participants use the different transportation modes can be made. The high probabilities for staying at each node, are the result of the coordinate by coordinate classification process. The relatively low transition probabilities of the wheelchair user reflect the fact that this individual uses the public transportation fewer times in the weekly sample of the analysis. This is contrasted with the crutches user that interacts with the public transportation network in a more regular way. In terms of frequency of transportation mode use, the crutches user is characterised by higher Walk/Rail transition probabilities, while the wheelchair user by higher Walk/Bus probabilities.

With regards to the overall strength of the travel mode preferences as expressed through the concentration parameters, the small (< 0.5) posterior value of the concentration parameter of the Dirichlet prior for the bus travel mode, particularly in relation to rail services of the crutches user reflects the lack of use of this particular mode. This update signifies redefinition of the LTDS derived assumptions, where the likelihood of bus use is greater. The uncertainty over the use of different public transportation modes of the wheelchair user is reflected by the increased overlap of the posterior concentration parameters. The exception is the walk state, which is in line with the prior assumptions for this user as expressed from LTDS.

Posterior inferences of external covariates has shown that their influence on the travel mode classification process varies between the two participants. In particular, proximity to available transportation modes and IMD had a reduced effect on participants interaction with the transportation modes, the magnitude of which is different between them. This magnitude varied between being statistically non significant and having a significant, but small effect. The former was the case for proximity to available transportation modes, while the latter was relevant for IMD. This could reflect the fact that the different use of transportation modes, as expressed by the different speed values, is not influenced significantly by the

chosen covariates. However, given the low accuracy of traces, this assertion should be verified by the use of more accurate position data.

In the context of the CAA modelling, the developed hierarchical DBN offers the advantage of providing information about the degree of interaction of individuals with the available transportation modes, together with the extensibility required to include a wide range of variables influencing it. Judging from the posterior densities of the concentration parameters, internal factors inherent to a person's capabilities, as expressed by the shape of the Dirichlet prior, play an important role in the interaction with the different transportation modes. On the other hand, external covariates have a limited effect on the inferred modes. The proposed approach is also able to characterise the transition dynamics of an individual through the use of the transition matrix.

6.3 Inferring activity types from unlabelled mobility data

This section's goal is to describe the DBN for activity type inference from unlabelled mobility data. In particular, section 6.3.1 defines the structure of the model, taking into account the potential activities that can be reached within a given isochrone polygon. Section 6.3.2 provides a detailed accuracy assessment of the degree of achievable accuracy under different model configurations using the social media dataset described in section 5.4. The methods and learnings are transferred in the context of the CAA model described in section 6.4.

6.3.1 Model specification

Similarly to section 6.2, the process of inferring activities from mobility data and additional evidence was formulated through a dynamic Bayesian Network.

The range of potential activities can be represented as the vector of potential destinations per activity category, as defined by the activities catchment area. In studies that involve activity type inference, and in particular the ones that use topic modelling methods, the nature of activity types at a destination can be captured by Points of Interest (POIs) (Gao et al. 2017). Approaching activity inference this way allows for a more direct quantification of uncertainty in activity estimates by assigning a probability at each potential activity depending on the absolute counts of potential destinations. Given the discrete nature of the set of activity events, this vector is assumed to follow a Multinomial distribution $z \sim Mult(p_{1...\kappa}, n)$ with parameters $p_{1...\kappa}$ being the activity probabilities with $\sum p_{1...\kappa} = 1$ and n being the total number of potential activities. Within a

Bayesian setting, the activity event probabilities can also be modelled as random variables following a distribution (prior distribution). The parameters of this distribution can then be used to encode any prior information that is assumed to influence the activity event probabilities. Such information can relate to the characteristics of mobility data, such as activity start time and activity duration. Due to distribution conjugacy, a natural choice of prior distribution for multinomial random variables is the Dirichlet distribution, defined by a concentration parameter $\alpha_{1...\kappa}$ with $\alpha_i > 0$. The concentration parameter vector controls the amount of probability mass assigned to an activity event before any potential destinations are observed.

6.3.1.1 Specifying the prior distribution

In the case of activity inference using POI data as feature vector, specifying an uninformative prior distribution would lead to the posterior activity estimates to be dominated by the likelihood derived from the POI data, reflecting the proportion of activities residing within an activities catchment area. For the Dirichlet distribution, this translates to setting the concentration parameter vector to an array of ones: $\alpha_{1...\kappa} = \mathbf{1}$ which results in drawing prior samples from a uniform distribution on the probability simplex. On the other hand, an informative prior using characteristics of the trajectory such as start time and duration will weight the simplex accordingly.

The effect of the Dirichlet prior on the Dirichlet/Multinomial posterior parameter estimates can be seen from the form of the posterior. The Dirichlet probability density function is:

$$f(\mathbf{p}|\alpha) = \frac{\Gamma(\sum_i^\kappa \alpha_i)}{\prod_i^\kappa \Gamma(\alpha_i)} \prod_i^\kappa p_i^{\alpha_i - 1} \tag{6.5}$$

The posterior then is the product of the prior with the data likelihood:

$$\begin{aligned} f(\mathbf{p}|Data) &= f(\mathbf{p}|\alpha) \prod_{y_i \in Data} f(y_i|p) \\ &\propto \prod_j^\kappa p_j^{\alpha_j - 1} \prod_{y_i \in Data} \prod_j^\kappa p_j^{y_i} \\ &= \prod_j^\kappa p_j^{\alpha_j - 1 + \sum_{y_i \in Data} y_i} \end{aligned} \tag{6.6}$$

where y_i is the POI vector in an activity isochrone polygon.

It follows from eq. 6.6, that the posterior is also Dirichlet distributed with the concentration parameters acting as pseudocounts, weighting the parameter

estimates towards the prior distribution, an effect that is referred to as "shrinkage" (Gelman et al. 2013). Using this property, the propensity of activity types at a particular point in time can be included in the model as a vector of probabilities, prior observing the Multinomial POI vector.

In the context of the Dirichlet/Multinomial model of this study, the shape of the Dirichlet concentration parameter vector α can be used as a means to adjust the specific activity type distribution throughout the set of trajectory locations for an individual. High α values (depending on the base measure) indicate that specific activity types are more likely to occur compared to the ones with relatively low values.

For this study, the concentration parameters were estimated from the full Foursquare dataset following an empirical Bayes approach. Within this approach, the prior distribution is estimated from the data and is considered an approximation to a complete hierarchical analysis where a probability distribution is placed on the prior distribution parameters (Gelman et al. 2013). In this way, posterior activity estimates for an individual are allowed to be influenced by the full Foursquare population level activity estimates. In a setting using a different dataset, such information can be obtained from supplementary data such as travel surveys as demonstrated in section 6.2.

Using the complete Foursquare dataset, a combination of checkin time and duration between subsequent checkins was used by calculating a gaussian kernel density estimate (KDE) for each activity type at each trajectory point and generating samples for each checkin/duration pair (figure 6.17). Note that the duration variable does not correspond to duration of stay, as this information is not available in the Foursquare dataset. Nevertheless, duration as calculated by the time elapsed between subsequent checkins has been used in studies using datasets with similar shortcomings, such as AFC (Lee & Hickman 2014, Alsger et al. 2018). The sampled values were organised in a vector for each activity type and each activity location. To ensure the concentration parameters follow an exponential distribution with rate proportional to the magnitude of KDE density for each check-in/duration pair and allow for more flexible priors, the resulting values were multiplied by Gamma distributed random variables with shape and rate parameters of the Gamma distribution $a = b = 1$.

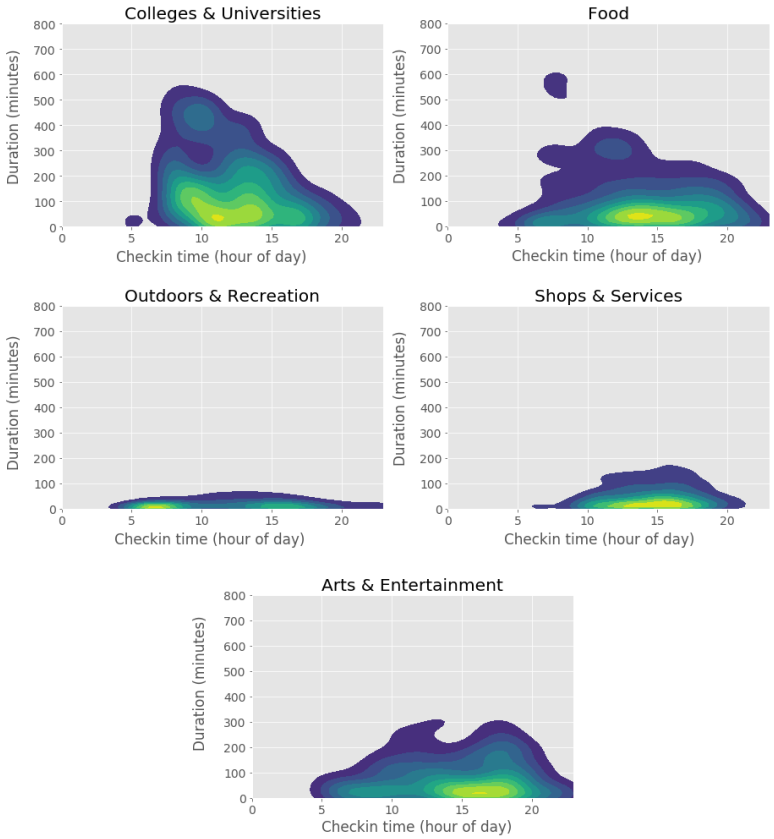


Figure 6.17: KDE contour plots of checkin time and duration between checkins for each activity type category

6.3.1.2 Specifying the dynamic component

In general, the sequence of activity types within an individual's trajectory are characterised by recurring patterns. Among other factors, this is the result of an individual's activity type scheduling processes (Kitamura et al. 1997). This property has been recognised by researchers as important for a number of reasons. First, it allows for a more realistic modelling of activity type patterns that is comparable with human decision making process (Allahviranloo & Recker 2013) and second, it enables more robust modelling, especially if the task is predictive inference. As already demonstrated in section 6.2, a particularly ubiquitous framework for modelling transition dynamics are Markov models. Markov models use the sequential nature of observations to estimate a transition probability matrix which can then be used to generate future model states. Specific examples within the task of activity modelling include the work of Allahviranloo & Recker (2013), where the parameters affecting activity sequencing were specified through the use of a Support Vector Machine model, while activity sequencing was modelled through

the use of CRFs. Other authors (Liao, Patterson, Fox & Kautz 2007) have used dynamic Bayesian Networks to model daily activity sequences given features such as previously inferred transportation mode and duration of trip segment as derived from GPS trajectory data.

A disadvantage of the use of Markov models in the context of many applications is their "memory-less" property. This specifies the conditional dependency of a future state with respect to the immediate previous one. For activity modelling this is a strong assumption as activities usually depend on temporal factors rather than the sequence that were carried out. For example, activities related to education is more likely to be dependent on the time of day rather than the nature of the previous activity. Nevertheless, the memory-less assumption has been widely adopted in the literature for trip purpose inference (Popkowski Leszczyc & Timmermans 2002, Han & Sohn 2016).

For this study, the dynamic component was modelled using a transition probability matrix with the rows specifying transition probabilities between different latent activity types. A transition matrix T of an K state Markov process is given by:

$$T = \begin{bmatrix} \pi(1,1), \pi(1,2), \ldots, \pi(1,K) \\ \pi(2,1), \pi(2,2), \ldots, \pi(2,K) \\ \vdots \\ \pi(K,1), \pi(K,2), \ldots, \pi(K,K) \end{bmatrix} \tag{6.7}$$

where each entry corresponds to the probability that the system transitions to state j given the state was i at the previous step:

$$\pi(i,j) = P(x_{t+1} = j | x_t = i) \tag{6.8}$$

The rows of the transition matrix were modelled using independent Dirichlet distributions with all concentration parameters equal to one, corresponding to no prior assumptions related to the sequence of activity types. This allows the resulting transition probabilities to be inferred only by the sequence of activity types while ensuring the rows of the transition matrix $p_i = \pi(m,i)$ is $0 \leq p_i \leq 1$ and $\sum p_i = 1$. This is a fairly common Bayesian approach when the transition probabilities are unknown or uncertain (Jaulmes et al. 2005, Bertuccelli & How 2008).

Consolidating the above, the model structure is illustrated in Figure 6.18. In this figure, the greyed square nodes represent observed data while circle nodes rep-

resent stochastic variables. The matrix notation represents the transition matrix between κ activity types.

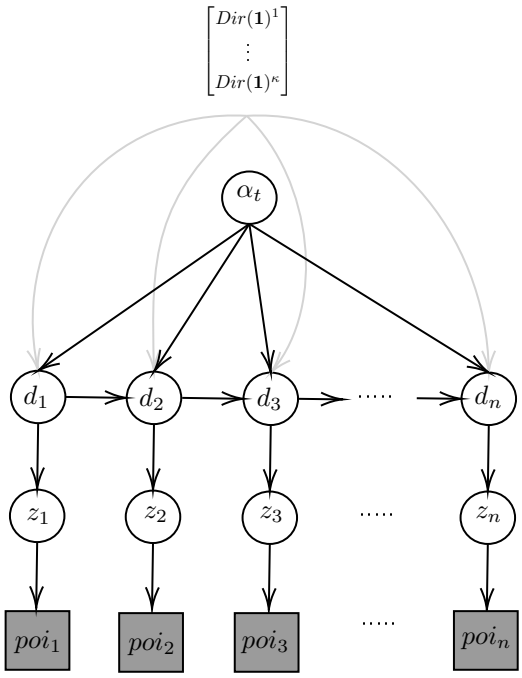


Figure 6.18: Graphical representation of the DBN used for activity inference.

More formally, the model is:

$$p(d_n|T, \alpha_{t_i}^{dur}) \sim Dir(\alpha_{t_i}^{dur}) \tag{6.9}$$

$$\text{with likelihood} \begin{cases} f(x, \alpha_{t_i}^{dur}) \\ \quad f(T_{row=argmax(d_{n-1})}, \alpha_{t_{i-1}}^{dur}), & \text{if } t_{i-1}\text{-}3<t_i<t_{i-1}+3 \text{ ,} \\ f(x, \alpha_{t_i}^{dur}), & \text{otherwise} \end{cases}$$

$$p(z_n|d_n) \sim Mult(poi_n, d_n)$$

For this study, a varying prior α_t^{dur} was introduced, that changes depending on the hour of day at a location i ($t_i = \{1...24\}$) and the time lapsed between subsequent locations (dur), specified in section 6.3.1.1.

The transition matrix was used to update the likelihood of the hidden activity sequence vector d under the assumption that an activity state is dependent on the previous state only if it falls within the same temporal window with the previous activity. This temporal window was specified to be +/- three hours from the check-in time to reflect plausible activity sequences among in the trajectory.

Table 6.6 below summarises the notation of the model:

Variable	**Description**
poi	POI- derived classification vector within an isochrone boundary
z	Multinomial probability distribution of activities.
d	Dirichlet distribution on z.
α	Concentration parameter vector derived from time of day and duration between check-ins.
T	$\kappa \times \kappa$ transition matrix.
t	Hour of day index.
i, n	Check-in index, Total number of check-ins per individual Foursquare user.
κ	Activity categories index.

Table 6.6: Description of variables of figure 6.18

6.3.2 Results

The model specification described in section 6.3.1 was applied to each individual Foursquare sequence of check-ins for each individual (trajectories). As has already been mentioned, this dataset is used within a completely unsupervised setting, using only the (weighted using labour demand data) Foursquare POI vector, activity sequence dynamics and check-in time/duration between check-ins as input to calculate activity type probability vector. Inference was performed using the well known Metropolis-Hastings sampling scheme described in detail in (Neal 2000).

A lognormal proposal distribution is used with the step scale modified in each iteration to increase the acceptance ratio. In the case of Dirichlet distribution, proposed values were normalised to sum to one to produce a valid proposal vector. For the rest of the model's nodes, an adaptive metropolis algorithm was used (Haario et al. 2001), with a scaled covariance matrix for the jump distribution to minimise the likelihood of invalid proposals. For each dataset corresponding to an individual's trajectory, two parallel MCMC chains were initiated with random starting values, for a total number of 10000 iterations. The first 1000 samples were discarded as not representative of the posterior distribution.

Next, the posterior quantities are presented for the inferred variables, having as benchmark POI vectors under the 5 minutes walking distance isochrone. Convergence of the MCMC chains was assessed using Geweke's diagnostic (Geweke et al. 1991). Stochastic variables that have z values within two standard deviation values around zero signify a MCMC chain that have converged. Figure 6.19 shows a plot with Geweke's z score for the inferred stochastic variables of the model for all participants.

As it can be seen, the bulk of the z scores lie within the boundaries of two standard deviations from zero. For the remaining variables, additional samples would have assisted convergence.

The posterior distribution of the latent parameter vector d_n corresponds to the probabilities of activities inside an isochrone polygon (in this case the 5-minute boundary). The results for each participant is shown in Appendix C. As it can be seen, the inferred concentration parameter vector α had a smoothing effect in the posterior quantities of activities for nearly all participants. In contrast, in the case of sparse data distributions (isochrone polygons having very few POIs) the posterior quantities are dominated by the empirical Bayes prior, as in the case for user #16 or #17 for example.

The posterior densities of the transition matrix capture the interactions of the users with the activities, provided that these occurred within the specified time framework set out by the model. Figure 6.20 below shows the posterior

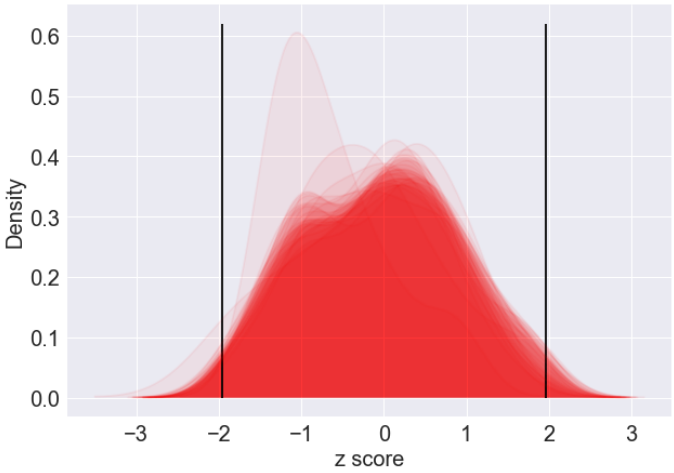


Figure 6.19: Geweke's z score for all stochastic variables

distributions for all users and for each element of the transition matrix.

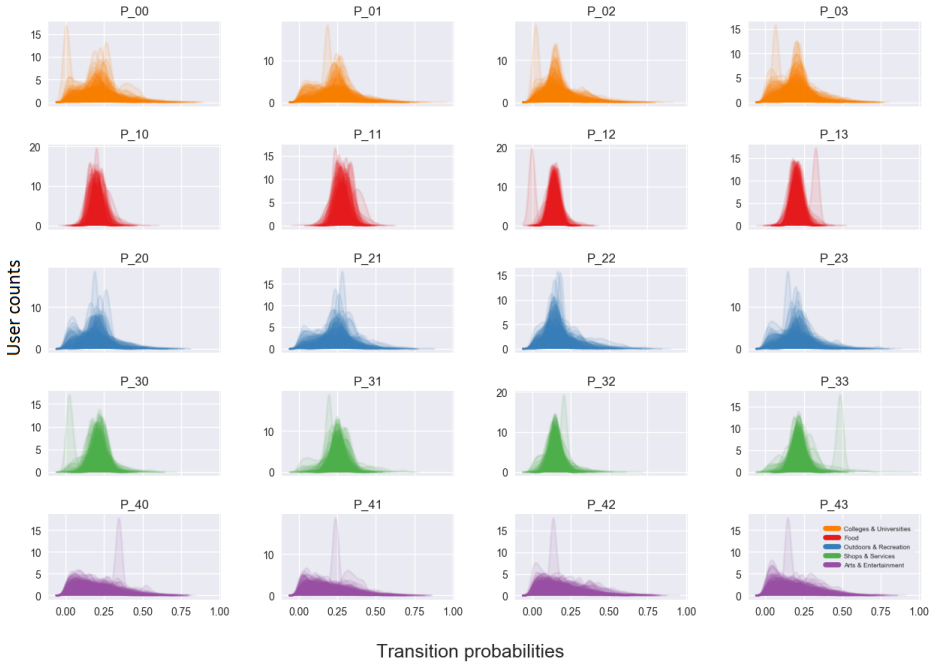


Figure 6.20: Transition matrix posterior distributions for all users. The figure titles $P_{00} \ldots P_{43}$ refer to transition probabilities between activity types. The transition probabilities of individual users for each transition are overlaid in each subfigure.

Modelling the transition dynamics between activities provides an additional insight to the activity patterns of individual users. In the case of this study, a bimodality can be observed between the interaction of activity *Colleges and Universities* with the rest of activities. This could potentially be attributed to

student and non-student population groups. The same is observed for *Outdoors and Recreation* activity which could signify users with different outdoor activity levels. It should be noted that the final column of the transition matrix can be derived in a deterministic way by $1 - \sum_j P_{ij}$ as the rows of the matrix must sum to one.

Finally, it should be noted that although the model provides interesting insights on the mobility patterns of individual users, interpretation of these patterns at an aggregated level should be done with caution and should not be regarded as representative for the whole population of Foursquare users. This is due to the limited number of trajectories used in this analysis.

6.3.2.1 Performance of activity type inference under different POI configurations

Using the self reported check-in activities of each individual user, the performance of the activity inference model could be evaluated under each different POI configurations corresponding to the different isochrone extents.

In addition to accuracy, two additional measures of performance were used: The AUROC (Area Under Receiving Operating Curve) and log-loss metric.

A ROC summarises the performance of a classification algorithm by representing the trade off between true positive (recall, sensitivity) $TPR = \frac{TP}{TP+FN}$ and false positive detection rate (1 - specificity) $FPR = \frac{FP}{FP+TN}$. Computing these two metrics for different thresholds and plotting these two quantities against each other yields a ROC.

By calculating the area under ROC (commonly by trapezoidal integration) one obtains the AUROC metric which ranges between 0-1, with 1 corresponding to perfect classification performance, and a value of 0.5 corresponding to a random classifier. This metric has several advantages over other metrics such as accuracy, since it is not sensitive to class distribution prior to inference and giving low scores to "one class only" classifiers (Bradley 1997). Moreover, it has an intuitive statistical interpretation, as it represents the probability that a randomly chosen false positive sample will have a lower probability than a randomly chosen false negative sample (Huang & Ling 2005).

It should be noted that the AUROC metric has traditionally been used within supervised classification settings, within which the classification algorithm is trained with a ground truth dataset. However, in this study, inference of the unknown model parameters was performed using the information contained in the POI vector and assumptions about user's activity patterns as included through the prior and the transition matrix. Nevertheless, since the encoding of the POI vector

is the same with the ground truth self reported check-in activities, it is possible to use this metrics to assess activity detection performance under the different configurations of POI vectors.

For the computation, the $argmax_{\kappa}(d_i)$ of the posterior distribution of the d_i random variable was taken for each user trajectory. Since AUROC is defined over binary classification frameworks, the activity classes were binarised with respect to each over, and the metric was computed for each individual activity class.

In addition, to assess the correspondence of the posterior probability vectors d_i of each isochrone bounding area with users's self-reported activities, the log-loss was computed, having as reference a degenerate distribution constructed by the ground truth check-in activities per isochrone area. Log-loss naturally quantifies the performance of a model whose output is a probability distribution. As a function, it is closely related to cross-entropy and Kullback-Leibler divergence in information theory and, in the case of binary output, is defined as $-ylog(p) + (1-y)log(1-p)$ where p is the predicted class probability. This formula can be extended to the multiclass case by summing over the separate losses for each class label $-\sum_{\kappa=1}^{K} y_{\kappa} log(p_{\kappa})$. A value of 0 indicates a perfect correspondence (no information loss) while larger values correspond to less correspondence between the two distributions.

6.3.2.2 Performance assessment using AUROC

Looking at the 5 minute isochrone Area Under Receiver Operating Curve (AUROC) values (figure 6.21), one could see that for the majority of Foursquare users, the model resulted in values fluctuating around 0.6 for all activity categories, particularly for *Food* and *Shop and Retail*. This is to be expected since *Food* and *Shop and Retail* categories were the dominating activity labels for the majority of the POIs included in the 5-minute isochrone area. For activity categories *Outdoors and recreation* and *Arts and Entertainment* there is a relatively high number of AUROC values fluctuating around 0.5 indicating that for such cases, the model's output is indistinguishable from a random classifier. Under closer examination, situations such as sparse POI vectors or class confounding POIs within an isochrone area seem to trigger this behaviour. For the cases where AUROC values are below 0.5, the model systematically miss-classified the correct activity for the particular isochrone area. This behaviour mostly occurs when a POI vector conflicts with the ground truth activity category by a large extent, together with repeated user visits to the problematic isochrone area within a trajectory. An example is a repeated user visit to an outdoor area that is within an isochrone polygon containing a disproportionate large number of Food POIs. In this case,

the classifier will repeatedly miss-classify the activity as a Food activity resulting in an AUROC value below 0.5. A similar behaviour can also occur in the presence of sparse POI vectors where the posterior distribution of activities is dominated by the empirical Bayes prior, which, for some individual activity check-ins, does not correspond to the ground truth.

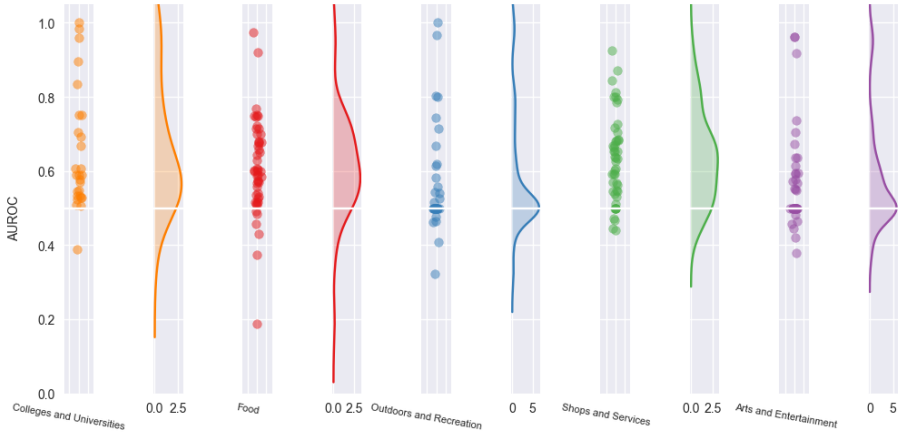


Figure 6.21: AUROC values for all user trajectories (5-min isochrone)

At a 10 minute walking distance isochrone, model's performance deteriorates for all activity categories, with more individual trajectories displaying systematic errors during activity inference. For some individual trajectories however, increasing the extent of the isochrone area seemed to have improved results for *Outdoors and Recreation*. This behaviour is most likely related to the more dispersed nature of Outdoor POIs within the 10-minute isochrone.

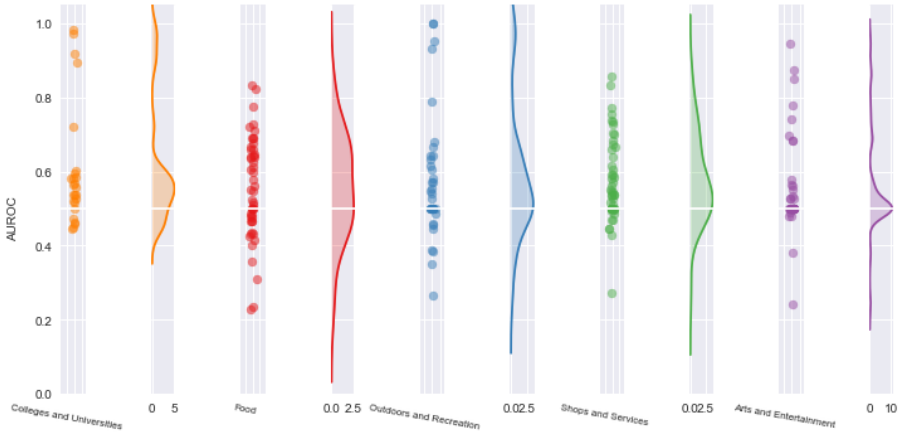


Figure 6.22: AUROC values for all user trajectories (10-min isochrone)

Further, at a 15 minute level isochrone, all activity inferences shrink further towards the AUROC 0.5 value, with increasing number of trajectories being systematically missclasified. At this level of resolution, most of the meaningful structure is lost from the data, resulting all activity categories behaving similarly. A similar situation occurs at the 20-min isochrone level.

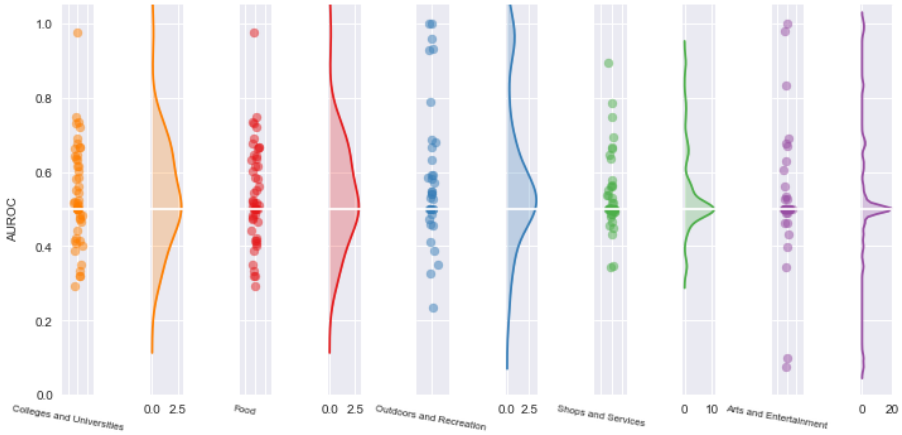


Figure 6.23: AUROC values for all user trajectories (15-min isochrone)

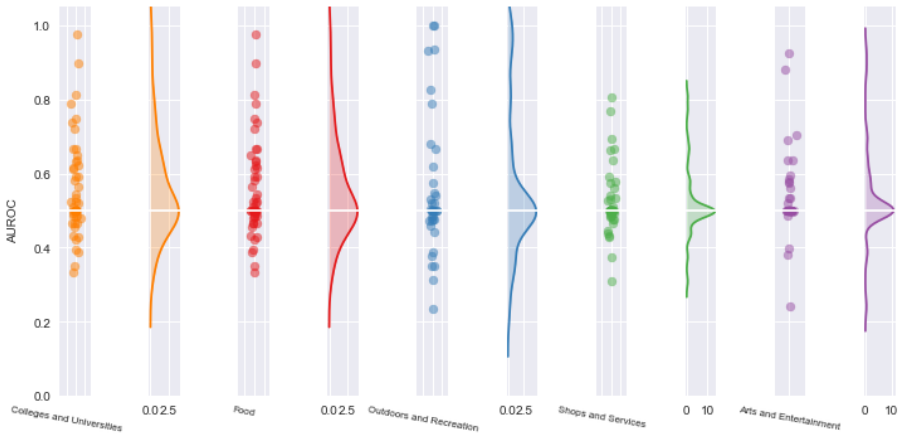


Figure 6.24: AUROC values for all user trajectories (20-min isochrone)

6.3.2.3 Performance assessment using log-loss

While the AUROC metric is useful technique to summarise the performance of a classification model under different data settings, it doesn't provide any insights on the performance of the classifier with respect to the probabilistic output of activities inference per icoschrone area. To address this, the log-loss was calculated, having as reference a (degenerate) distribution constructed using the ground truth check-in activities. To avoid numerical errors, a small jitter of the order of 10^{-3} was added to the reference distribution. A log-loss value of 0 assumes no information loss between the two distribution while increasing values indicate increased information loss.

Looking at log-loss values for the 5-minute isochrone (Figure 6.25) one could see that the majority of participant trajectories lie bellow a mean log-loss value of around 1.4, which translates to probability of 0.246 ($e^{-1.4}$) per each isochrone area, an improvement over a random guess for the 5 activity categories of this case study. Looking at the 1st and 3rd quartile spread of log-loss values, one could see that activity predictability varies greatly between and within users, indicating that the limits of activity predictability is both user and location dependent. The log-loss values gradually increase with increased isochrone bands, signifying gradual deterioration of results.

6.3.2.4 Accuracy assessment

Using the revealed activity types reported by the Foursquare users, the absolute accuracy for each individual can be calculated. Figure 6.26a shows the histogram of accuracy values for each individual trajectory and each isochrone band. For the 5 minute isochrone one observes a bimodality in the individual accuracies,

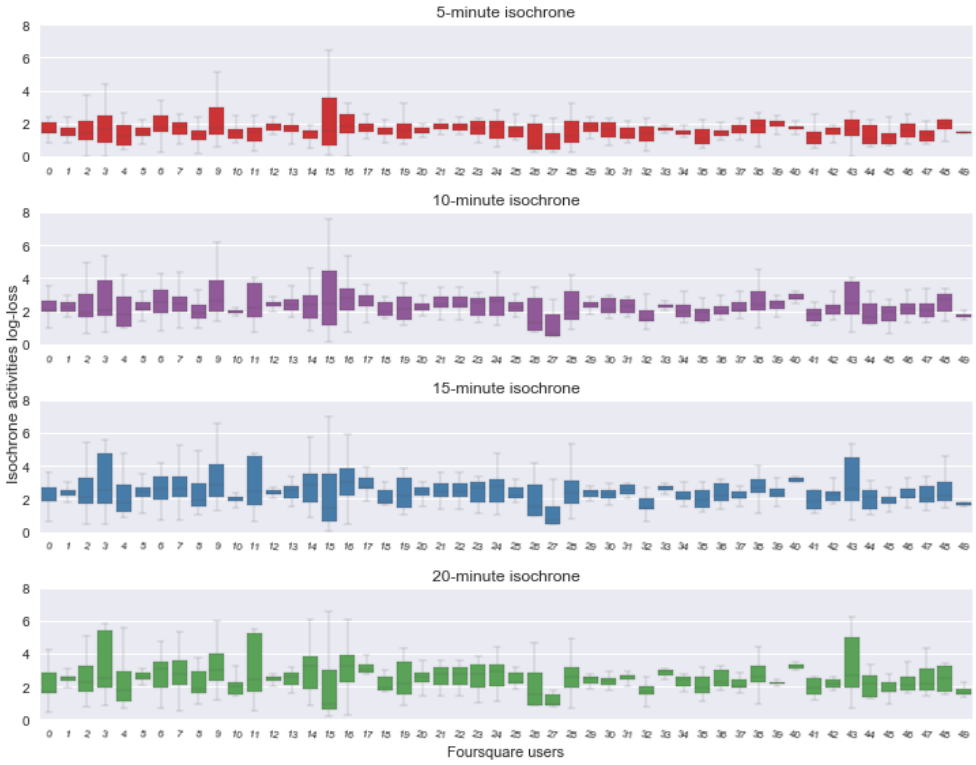
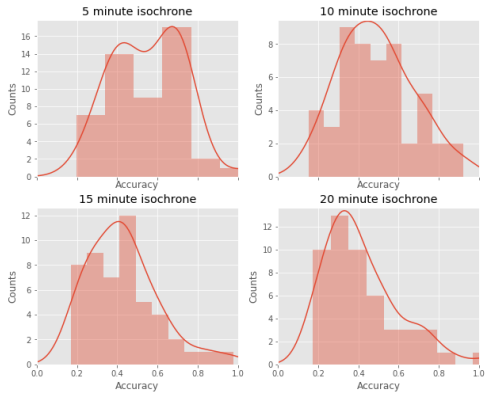


Figure 6.25: Log-loss boxplots of all user trajectories

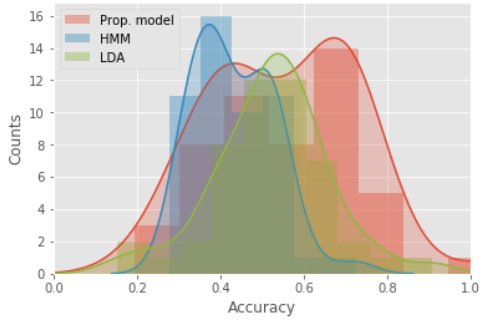
the first mode peaking around 0.4 while the second around 0.7. Under closer examination, trajectories with few data points and very sparse POI vectors seem to result in lower accuracy, as activity type inference is dominated by the checkin time/duration prior.

Finally, the model's output for the 5-minute isochrone is compared against two popular generative models in activity type inference: a hidden markov model (HMM) with Gaussian emission probabilities for check-in time and duration between subsequent check-ins, and a LDA model using the 5 minute isochrone POI vector as words, the individual check-in locations as documents and the activity types as (latent) topics (figure 6.26b). Contrary to the proposed method where the output probabilities are assigned to the corresponding POI activity type categories through the multinomial distributed POI vector, the labelling of the output of HMM and LDA correspond to activity type clusters and as such it requires an extra step of interpretation to assign semantic properties. In this study, this was done by attributing the ground truth activity type to the corresponding label according to a majority count of labels belonging to the particular activity type. As it can be seen, overall the proposed method resulted in increased accuracy compared to the HMM. The LDA model performed reasonably well, however

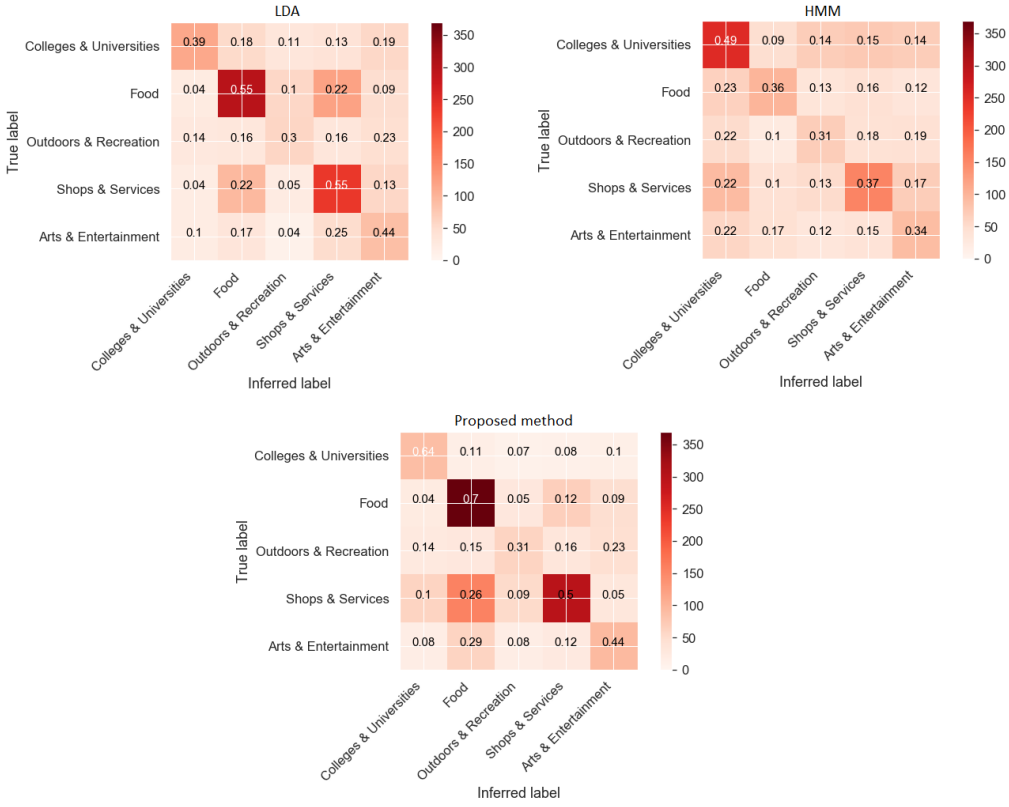
it lacked the flexibility to provide accuracy results for the individual trajectories that have that potential compared to the proposed method. The mean accuracy values for the three methods were 0.43, 0.52 and 0.56 for the HMM, LDA and proposed method respectively. Finally, the confusion matrices for all three models are presented in figure 6.26c.



(a) Histograms of individual accuracy values for the proposed method.



(b) Proposed method, HMM and LDA accuracy values.



(c) Proposed method, HMM and LDA confusion matrices .

Figure 6.26: Accuracy histograms and confusion matrices for the proposed method, HMM and LDA.

6.3.3 Discussion

The DBN used in this case study allowed for diffused estimates as the informed Dirichlet prior acts as a smoothing agent introducing pseudocounts to the Multinomial model through the Dirichlet/Multinomial conjugate property. The degree of smoothing depends on the prior information introduced by the Dirichlet concentration parameter. In the absence of data that could relate to personal characteristics as in the modelling approach of section 6.2, this case study used an empirical Bayes Dirichlet prior as an agent to incorporate population level characteristics dusting inference. Similarly to section 6.2, individual user's activity transition dynamics were modelled through a Markov specification using a stochastic transition matrix, allowing the extraction of characteristic activity profiles for each individual user.

The choice of isochrone levels was based on common assumptions related to maximum walking distance an individual is willing to traverse from a point of access (Transport for London 2010), and thus it allows the findings to generalise to other mobility datasets such as Automatic Fare Collection systems and cell-tower mobility data within dense urban settings. The former is of particular importance considering the fact that the following chapters use a lower resolution smart card dataset.

The performance of the model under the different isochrone configurations was assessed using the AUROC and log-loss metrics. Results have shown that activity detection benefits most from the 5-minute isochrone, however the 10-minute isochrone retains its integrity for most individual trajectories, particularly for categories such as *Colleges and Education* and *Outdoors and Recreation*. Larger isochrones yield inferior activity detection results triggered by systematic errors in the data and the lack of within activity class structure that can be exploited from the model as determined by AUROC values. The overall accuracy of the 5/10-minute isochrone activity type inference seem to be in par with other relevant studies (e.g. Allahviranloo & Recker 2015, Shen & Stopher 2013, Allahviranloo & Recker 2013) for non home/work related activities. However, the current study benefits from being validated in an unsupervised classification setting using revealed individual activity types, as opposed to proxy ground truth data such as travel surveys or synthetic data (e.g. Yin et al. 2017, Hasan & Ukkusuri 2017).

Similarly to the approach of section 6.2, limitations of this modelling approach can be found in the the computational intensive nature of the MCMC simulations which makes this framework not suitable for real-time applications. Moreover, it is unclear how this model will perform for inferring activities other than the ones that can be solely determined by characteristics of the build environment such as employment and home. It is speculated though that the modular structure of

this framework would be able to account for this challenge, either by modifying the likelihood function or by specifying an informed prior to incorporate added information such as duration of stay or sociodemographic characteristics. This is explored further in the following section.

6.4 A Capabilities Approach to accessibility: Model formulation

So far, sections 6.2 and 6.3 have shown how some of the core components of accessibility studies (interaction with transportation modes and interaction with available activities) can be extracted from unlabelled mobility data using DBNs. This section's goal is to consolidate the methodological approaches introduced in chapters 6.2 and 6.3 and formulate a Capabilities Approach to Accessibility (CAA) model, expressing the relationship between the accessibility components using the Capabilities Approach (CA). Following the schematic 3.1 of section 3.5.1 schematic 6.27 below presents a high level view of the approach.

Specifically, at the top of the hierarchy the functioning vectors consists of the personal characteristics, which are assumed to be individual dependent and persistent throughout an individual's trajectory. These are informed using travel survey data and define the prior assumptions within a Bayesian setting (see section 6.2). The environmental characteristics on the other hand, are assumed to vary throughout an individual's trajectory, depending on the location. This dynamic component is captured through the use of transition matrices between activity types and transportation modes, the rows of which are modelled using the environmental characteristics. Both personal and environmental characteristics inform the latent capabilities, which are characterised by a) the potential destinations/opportunities that are available to a person and b) the potential public transportation modes an individual can use to access them.

At the bottom of the hierarchy are the realised functionings, which are used to infer the capabilities. These are a) the available opportunities bounded by the activity space for each destination and b) the transportation modes used by the individual. This process is relevant for each individual and is dynamic, in the sense that the results in each time step influence the results of the next time step in the trajectory. Further implementation details are given in section 6.4.1 within the context of exploring the link between transport disadvantage and social exclusion in London. The data input for this model are individual trajectories from London's AFC system (Oyster card), however, as demonstrated in sections 6.2 and 6.3 the overall methodological framework developed (figure 6.1) is able to account

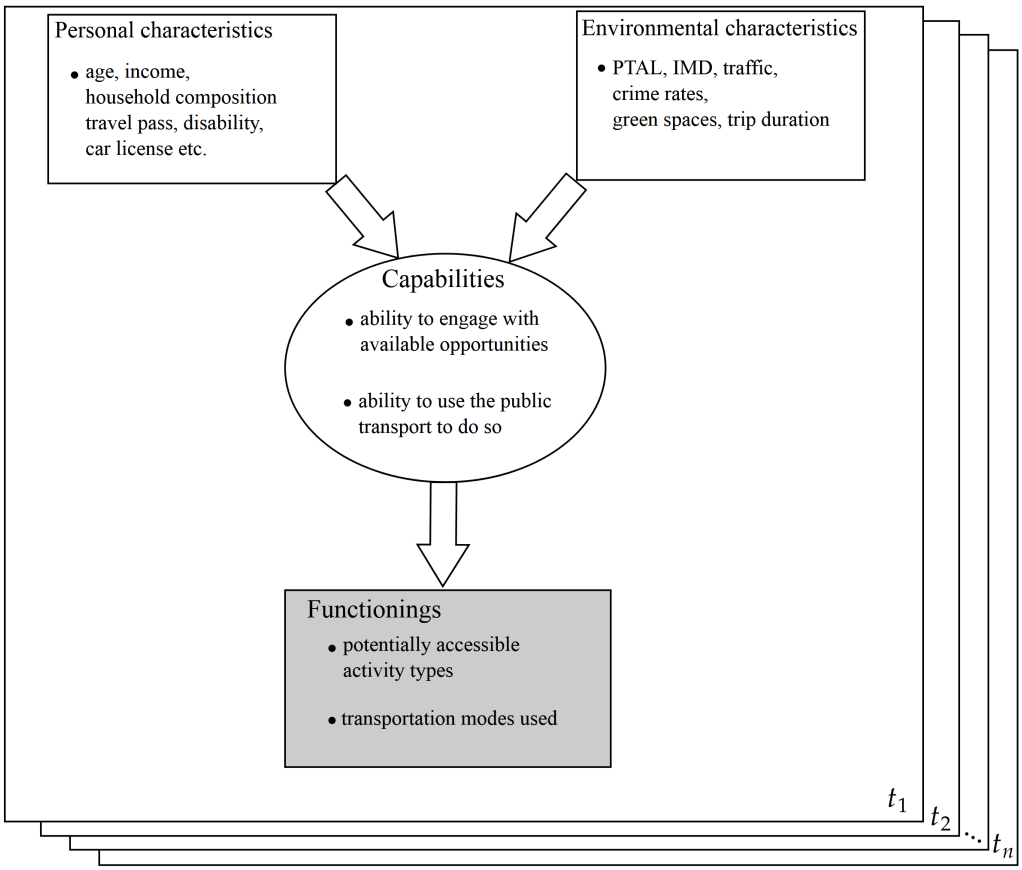


Figure 6.27: Roadmap for developing the Capabilities Approach to Accessibility (CAA) model.

for diverse sources of unlabelled data (such as CDR, GPS, social media data etc.).

The study focuses on three population groups: individuals having annual household income below £15000, individuals $>$ 60 years of age and an unconstrained base population group. The choice of those population groups was based on two factors: the sample size of each group and past research providing evidence of population groups with significantly different accessibility levels compared to the majority of the population (Páez et al. 2010, Hickman et al. 2017, Kamruzzaman et al. 2016, Titheridge et al. 2009).

6.4.1 Model specification

In the specification of this case study, two distinct but interacting components of an individual's act of reaching opportunities are included:

- Ability to interact with the available public transportation modes
- Ability to interact with the available destinations/opportunities

Similarly to sections 6.2 and 6.3, the relationship between these components as well as the personal characteristics and external factors were modelled using dynamic Bayesian networks. This time however, the link between capabilities and functionings is made explicit by providing a definition of the elements of the capabilities set and how these relate to the observed functionings.

6.4.1.1 Inclusion of personal characteristics

Although it is generally agreed that the personal characteristics of an individual influence mobility to a large extent, it has been argued that such variables express the potential of individuals to travel and as such they can not be used on their own to characterise accessibility (Kamruzzaman et al. 2016). Moreover, as these factors tend not to change over large periods of time, they cannot be used to describe the day to day interaction with the available transportation means and activity types. Nevertheless, personal characteristics provide the core information for many accessibility related studies and as such they have been used in conjunction with other datasets for identifying social exclusion (Casas 2007), defining the extent of activity spaces (Li & Tong 2016) and measuring access to activities and public transport (Wixey et al. 2005).

In this context, personal characteristics is an overarching term including the demographic characteristics and potential resources that could benefit or impede accessibility of an individual. Throughout the literature, variables such as ethnicity, age, gender, health, occupation status as well as income and access to car (Hananel & Berechman 2016, Preston & Rajé 2007, Simma & Axhausen 2003) have been found to be some of the most significant explanatory variables of an individual's access to destinations.

For this case study, two different data sources were used:

- LTDS survey described in section 5.5 for interacting with the available public transportation modes
- London's Rolling Origin Destination Survey (RODS) for interacting with the available activity types at a destination

Inclusion of personal characteristics in the model was based on the specification of section 6.2. Sections 6.4.1.1 and 6.4.1.1 describe the procedure for both transportation mode and activity types.

Transportation mode modelling

Contrary to section 6.2 where the vector of personal characteristics was limited to

age and disability status, the direct link between Oyster card records and LTDS allows the extension of the factors influencing an individual's interaction with the public transport modes to include a wider range of sociodemographic characteristics and resources. Table 6.7 below provides a description of the variables considered:

Variable	Cardinality	Description
Age	Continuous	Age of the respondent
Income	11	Gross household income before tax
Household members	Continuous	Number of household members
Travel pass	4	Whether the respondent possesses a travel pass
Disability	9	Disability type
Car license	2	Holds a full driving license
Occupation	9	Main occupation
Free pass	2	Whether the respondent hold a free travel pass
Sex	2	Sex of the respondent
Working status	12	Respondent's working status
Ethnic group	19	Respondent's ethnic group
Health problem	2/2	Long/short term health problem
Car use as driver	8	Frequency of car use
Car use as a passenger	8	Frequency of car use as a passenger
Regular taxi use	8	Frequency of black cab use
Private taxi use	8	Frequency of use of minicab, Dial a ride etc.
Walking	8	Walking frequency

Table 6.7: LTDS personal characteristics used in the case study

Similarly to section 6.2.1, a set of individual binary logistic regression models were fitted for each transportation mode $\kappa \in \{Bus, Rail, Tram\}$ category to compute the relative frequency of an individual using a particular transportation mode given the set of covariates/factors of table 6.7. The choice of individual binary logistic regression models over a multinomial regression model was based on the absence of mutually exclusive transportation choices over the set of available transportation means, as well as the absence of category specific covariate/factor information. Similarly to section 6.2.1, the breakpoint to coding the response variables was the frequency of use of a particular transportation mode (more than once per month).

The results of the logistic regression analysis using the updated variable vector can be found in Appendix A. In all cases, the baseline category was the outcome of not using the particular transportation mode frequently, while the log odds ratio is defined as the logarithm of the ratio between the two outcomes:

$$log(\frac{P(Y=1)}{P(Y=0)}) = \beta \mathbf{X} \tag{6.10}$$

where β is the vector of regression coefficients and $\mathbf{X}$ is the design matrix of covariates/factors.

Exponentiation of the coefficients β in the above equation, allows the recovery of the relative influence of a particular variable expressed as the odds-ratio of a unit increase in X_i for $i \in \{1...\#variables\}$.

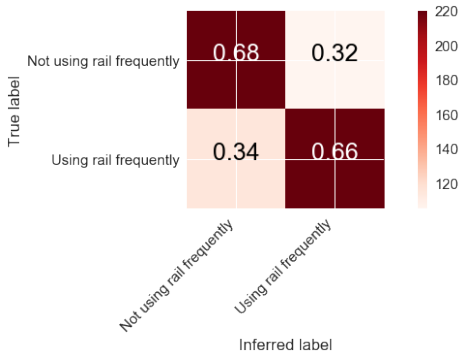
In the case of travelling by bus, disability, income, working status, possession of a travel card as well as car were found to have a statistically significant effect in determining the use of bus over the rest of the transportation modes. Higher incomes ($> 75000£$ per year) have a decreasing influence of using the bus over the rest of the modes, with odds ratio of 0.5. Low income statuses such as being a student, being unemployed or the inability to work because of a health problem increase the odds of choosing the bus by a factor ranging between 1.7-2. In terms of commodities, possession of a travel card increases the odds of using the bus, while access to a car decreases the odds by 0.84.

Examining the results of the logistic regression for rail, disability, age, income, sex, employment status, type of profession and ethnic group are among the sociodemographic variables that have a statistically significant effect on the odds of choosing a rail over the bus and tram. Similarly to bus, mobility impairment has a decreasing effect ranging between 0.5-0.6. Besides mobility impairment, cognitive impairment also has a decreasing effect in using rail over the rest of the transportation modes. Age has a marginal, but statistically significant decreasing

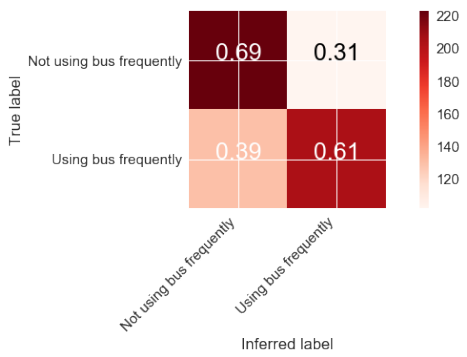
effect on rail frequency of use (0.97). Income on the other hand has an increasing effect for a wide range of income strata over 50000£. Participant's gender was also found to be statistically significant, increasing the odds of using rail modes by 1.45 in the case of male. Similar to the results for bus, being unemployed, retired, working from home or volunteering decreases the odds of using rail over the rest of the modes. Finally, ethnicity was found to have a significant effect with the factor *Other White (English/Welsh/Scottish/Northern Irish)* decreasing the odds by 0.54 and *Arab* increasing the odds by 1.81.

Results for choosing tram over bus and rail revealed a statistically significant effect for the variables *household members, car license, free travel pass* as well as *ethnic group*. In particular, the odds of choosing tram over the rest of transportation modes increases marginally by a factor of 1.13 with increasing household members. On the other hand, possession of a car and absence of free pass decreases the odds by a factor of 0.56-0.68. Ethnicity was found to have strong increasing effect for a wide range of ethnic groups (*Chinese, Indian, African and Caribbean, Mixed or multiple ethnic groups*) ranging from a factor of 8-13 when determining the odds of tram frequency of use.

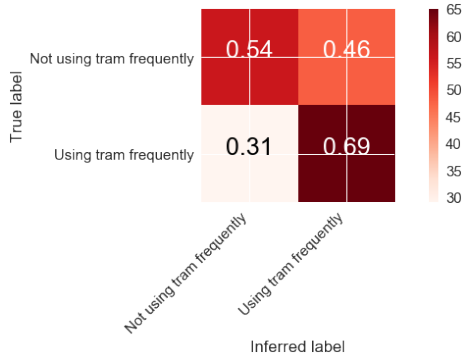
Figure 6.28 below shows the results of the individual logistic regressions on predicting the transportation mode given sociodemographic characteristics on the test dataset after performing a 70-30% train-test split on LTDS data.



(a) Confusion matrix for rail use



(b) Confusion matrix for bus use



(c) Confusion matrix for tram use

Figure 6.28: Confusion matrices for the individual LTDS regressions

Destination purpose modelling

The influence of personal characteristics for the task of destination inference was determined using the RODS dataset. The RODS survey is a recurring yearly questionnaire survey conducted at TfL's underground stations for the purposes of capturing information about London underground users' journeys. The dataset contains a multitude of variables such as: Line loading by section/line and time of day, route choice by origin-destination pair, numbers and types of interchanges, estimated and expected journey time, destination shop/activity as well as personal characteristics such as sex, age and disability etc. RODS distinguishes a journey's destination purpose by the types specified in table 6.8. The total number of the questionnaire respondents was 31854, with the majority (around 84%) of the destination type responses being *Travelling to work*.

To be consistent with the OS POI categories RODS destination purposes were re-categorised according to the categories of OS POI dataset. For the category Eating & Drinking, the destination names (in the case they existed) were matched to the corresponding names in the OS POI dataset. For the rest of the destination categories, the re-categorisation was based on table 6.8.

RODS destination categories	**OS groupings**
Shopping	Retail
Theatre /cinema /concert / Sporting activity /event	Outdoors and Recreation
Museum /exhibition	Outdoors and Recreation
School /college /university (as student)	Education and Health
Personal business (e.g. doctor/dentist)	Education and Health
Normal workplace /Other workplace /employers business	Work

Table 6.8: Re-categorising RODS destination purposes

Contrary to the methodology followed for determining the influence of personal characteristics on transportation mode described in the previous paragraph (**Transportation mode modelling**), the range of destination categories was modelled using multinomial logistic regression with *work* being the baseline category. This was due to the lack of overlap between different destinations, as the RODS dataset contains one destination purpose per record. The frequency of observations in the dataset is heavily skewed towards the *work* category, accounting for nearly 83% of the total records. As a result, prior to the implementation of the regression model, *work* category was randomly downsampled to match the relative frequency of the remaining destination categories. The final dataset included 200 samples for each destination category, 30 % of which was kept for testing purposes. The variables used for predictors were chosen to be disability, age, sex and self-reported arrival time.

Looking at the regression results (Appendix B), self-reported arrival time was found to have a statistically significant effect in predicting the destination purpose for all categories relative to the baseline category (*Employment*). In all three cases, the odds-ratio is > 1 indicating that later hours of day increase the odds of predicting categories other than the baseline, with *Eating and Drinking* and *Sports and Entertainment* having the largest odds ratio. Age was found to have a significant effect as well for the categories *Education and Health* and *Outdoors and Recreation*. For the category *Education and Health* the odds ratio was found to be < 1 indicating that increased age decreases the odds of predicting this destination

Variable	Cardinality	Description
Disability	2	Whether the respondent has reported a disability that limits their journey
Sex	3	Male/Female/not answered
Age	Continuous	Age if the respondent
Arrival time	Continuous	Estimated arrival time of the respondent

Table 6.9: RODS personal characteristics used in the case study

category over the baseline. This is not the case for *Outdoors and Recreation* where the odds ratio was found to be 1.34. Finally, the factor disability was found to have a significant effect for *Sports and Entertainment*, increasing the odds for predicting this category over the baseline category nearly four times when a person has not reported a disability. Besides time of day, the covariates used did not have any statistically significant effect in predicting the categories *Eating and Drinking* and *Retail*, a fact which manifested itself in reduced predictive accuracy for these categories. This can be attributed to the increased class overlap between these two activity categories, as well as the reduced precision of *Retail* and reduced recall for *Eating and Drinking* category. Figure 6.29 shows the results of the Multinomial regression on the test dataset derived from the 70-30 % train-test split process.

Similarly to the LTDS dataset, the RODS multinomial logistic model was used to predict the most likely destination purpose from the Oyster card/LTDS dataset.

6.4.1.2 Including external factors

In the scope of the CAA, external factors refer to environment variables that could influence the choice of an activity type within an activity space, or variables related to availability/reliability of public transport services. Similarly to chapter 6.2, these variables are assumed to change as the individual moves through space.

Different data sources were used for this task in this case study. This included census statistics, data from UK's Department of Transport, the London Metropolitan Police Service and OpenStreetMap as well as data derived from an individual's trajectory (such as trip duration). Similarly to section 6.4.1.1 these are presented both for transport mode modelling and activity type modelling.

Transportation mode modelling

External variables influencing an individual's ability to use the public transport

Figure 6.29: Confusion matrix of RODS multinomial regression

modes can be diverse and multi-faceted, ranging from the transportation network attributes, such as connectivity, to characteristics of the wider physical and social environment such as safety/security and street lighting (Wixey et al. 2005). For this case study, two variables were included: Public Transport Accessibility Levels (PTAL) and trip duration.

London's PTAL Public transport accessibility level (PTAL) is a measure of service availability, used in a variety of contexts ranging from transport and urban planning (Evans 2009) to equity in transport provision (Wu & Hine 2003).

London's PTAL was first designed by the borough of Hammersmith and Fulham in 1992, and after a series of reviews and tests, it has been agreed by the borough-led PTAL development group as the most appropriate method for calculating accessibility to public transport in London (Transport for London 2010). The index is particularly attractive in public transport appraisals since it combines different measures of transport quality of service such as:

- Walking time from any geographical point to the public transport access points

- Reliability of available modes of transport
- Number of services available within the catchment geographical area
- Average waiting time

One the other hand, its calculation does not take into account:

- The speed or utility of accessible services
- Crowding, congestion or the ability to board a service in general
- The ease of interchange

Calculation of PTAL is relatively easy and straightforward. The reader is referred to Transport for London (2010) for details. The final output is 6 levels of accessibility covering the Greater London Area ranging from low to high (0-6). Levels 1 and 6 are further subdivided for clarity.

As it can be seen from figure 6.30, PTAL essentially quantifies the density of public transport in an area. For the purposes of the analysis, PTAL levels were given a weight from 1 to 9 (lowest to highest including the subdivision of classes 1 and 6).

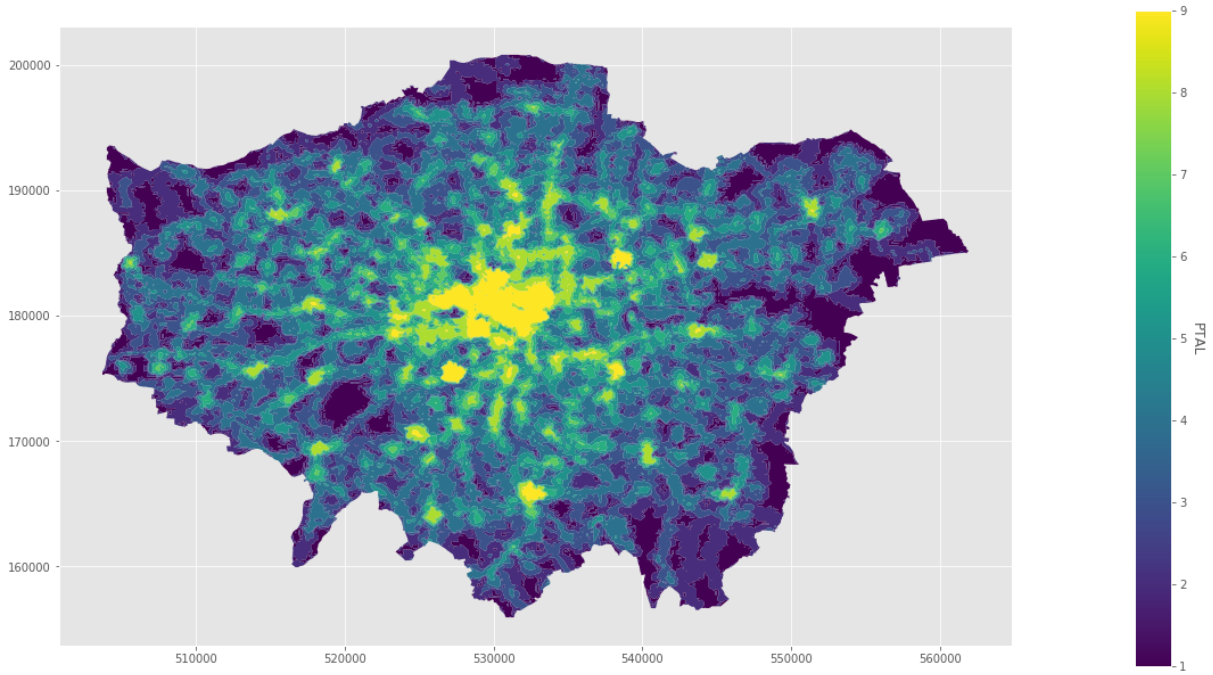


Figure 6.30: London PTAL

Trip duration Trip duration is the second external variable used in modelling an individual's transition between public transport modes. This variable has been widely used within different accessibility contexts ranging from space-time prism approaches (Kwan 2013) to utility based accessibility measures (Geurs et al. 2010) as it is believed to be one of the most important infrastructure based accessibility indicators.

Having reconstructed the individual trip segments for rail, bus and tram journeys (see section 5.5.3.1) it is possible to calculate the duration of a trip segment. For tube services this is straightforward as it can directly be derived from transaction times during a 'tap' in and out. However, for bus and tram services, where the alighting stop/station is inferred, trip duration has to be estimated. For this case study, individual trip duration and total travel time of the bus/tram trips was computed by placing requests to TfL's unified API journey planner with the following configuring parameters:

- Date of travel: Taken from the Oyster card transaction time
- Travel mode: Bus only/Tram only
- Max walking time: None

TfL's journey planner API response is a list of travel options matching the query criteria for a particular origin/destination pair. From these potential routes, only the ones that contained the bus route number as appearing in the Oyster card/LTDS dataset were selected. The mean travel time was then computed from all the candidate bus routes/tram services, and assigned as trip duration for the particular trip segment.

Destination purpose modelling

External variables associated with the destinations where the different activity types are located can either support or impede the choice of an activity. From the multitude of variables referenced in the literature (such as street furniture, width of street etc.) (Evans 2009), deprivation, proportion of green areas, traffic and crime stand out as being important factors influencing the choice of activity.

Index of multiple deprivation (IMD) The level of deprivation of an area is considered to be an important variable contributing to transport related social exclusion (Casas et al. 2009, Kamruzzaman et al. 2016) as well as the ability of an individual to participate in social and economic opportunities (Titheridge et al. 2009).

National statistical agencies regularly collect such datasets, and for the UK there exists theIndex of Multiple Deprivation (IMD), which assigns a composite score to an area derived from data related to income, employment, education, health, crime, housing and environmental quality (McLennan et al. 2011).

The advantage of using such an index is that it greatly simplifies the analysis as it allows the indirect inclusion of different characteristics of the environment through a unified index. However, it assumes that all individuals respond to the different deprivation components in a similar way when reaching for a particular activity. This is a strong assumption as people with different capabilities might weight the relative influence of these components in a different way.

For this case study, data from 2011 were used as they are thought to be more representative of the 2013 Oyster/LTDS dataset. Figure 6.31 below shows the IMD values in London at lower super output area (LSOA) level, which is a geographic area used for small area statistics in England and Wales. The data were standardised to be within the 0-1 range.

Crime rates Crime has been identified by many authors as a negative factor influencing accessibility (Pyrialakou et al. 2016, Evans 2009). This can vary for different individuals depending on personal characteristics such as gender and age (Church et al. 2000, Schmöcker et al. 2008). For this study, London's metropolitan police API was used to retrieve all reported crimes [1] for a period of four years (2014-2017). The data were aggregated at the LSOA level. The dataset was standardised by the population density for each LSOA to give an indication of the crime rate, and min/max scaled to range between 0 and 1 for computational convenience. Figure 6.32 below shows the resulting dataset.

[1]The crime categories considered were: Anti-social behaviour, Criminal damage and arson, Drugs,Other theft, Violence and sexual offences, Other crime,Burglary, Vehicle crime,Public order, Shoplifting, Theft from the person, Possession of weapons, Robbery,Bicycle theft.

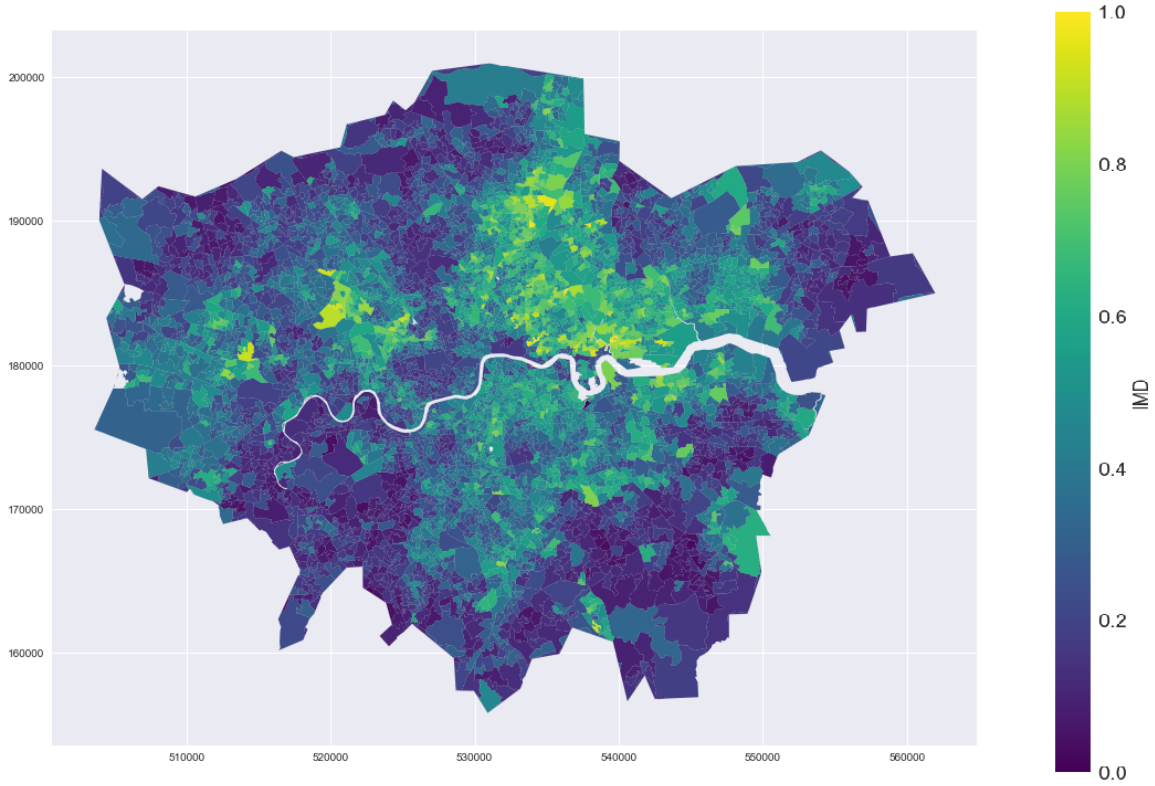


Figure 6.31: London IMD

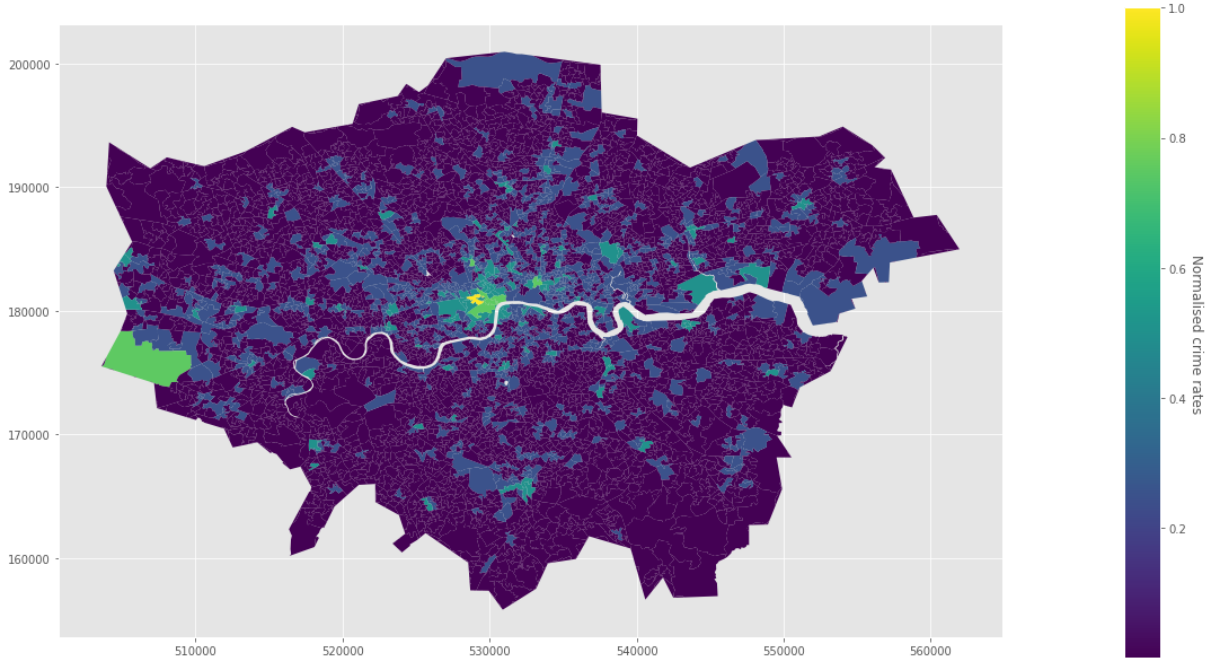


Figure 6.32: Standardised crime rates

Proportion of green areas Urban design elements have been argued to contribute to the confidence, movement and transportation choice for individuals belonging to specific population groups, particularly the elderly and disabled (Evans 2009). These design elements can be anything that affects the quality of the built environment such as railings, signs of vandalism and street furniture. For this study, the fraction of green areas was computed as a proportion of the total LSOA area as it has been shown to affect positively the aesthetics and perceptions of people for urban areas (Smardon 1988). The green areas were taken from OSM API using the tags that refer to the presence of greenery[2]. Figure 6.33 below shows the resulting dataset. As before, the values were min/max scaled to be within the 0-1 interval[3].

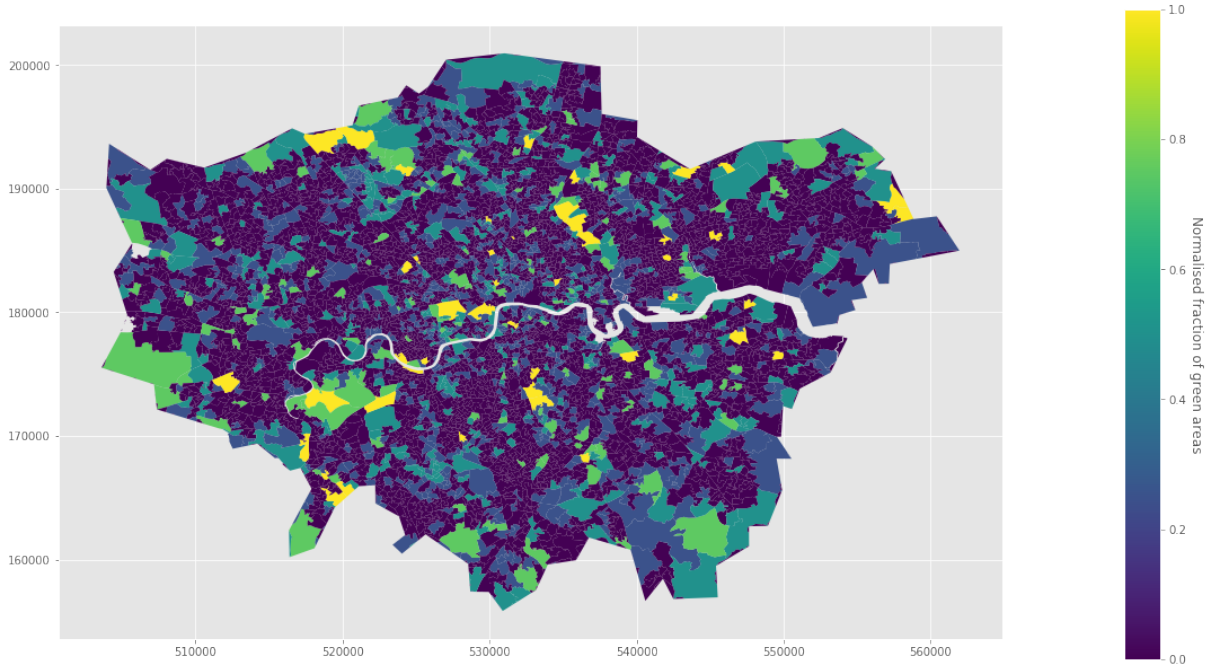


Figure 6.33: Standardised green space areas

[2]The OSM tags used were: leisure, park, garden, dog_park, allotments, forest, grass, greenfield, meadow, orchard, recreation_ground, village_green, wood, fell, grassland, heath, scrub, wood

[3]For this study, the scaling reflects the assumption that the proportion of greenspace in an area doesn't have a negative seasonal or diurnal impact to the ability of people to reach their activities, e.g. when it is dark or when it is winter

Traffic volume The fourth covariate considered is traffic counts for the Greater London Area. The volume of traffic is considered to have a negative impact on accessibility as higher volumes are thought to cause feelings of danger of travel, threat, anxiety, insecurity and stress (Evans 2009). Areas with high traffic volume have also been found to correlate with deprived areas, resulting in increased traffic related deaths compared to more affluent areas (DfT 2014). For this study, the traffic counts were used for the motorways and primary roads for the Greater London Area for all motorised vehicles, sourced by UK's department for transport[4]. For reasons of consistency, the traffic counts were re-aggregated at an LSOA level, standardised by population density and min/max scaled to be within the 0-1 range. Figure 6.34 below shows the resulting dataset.

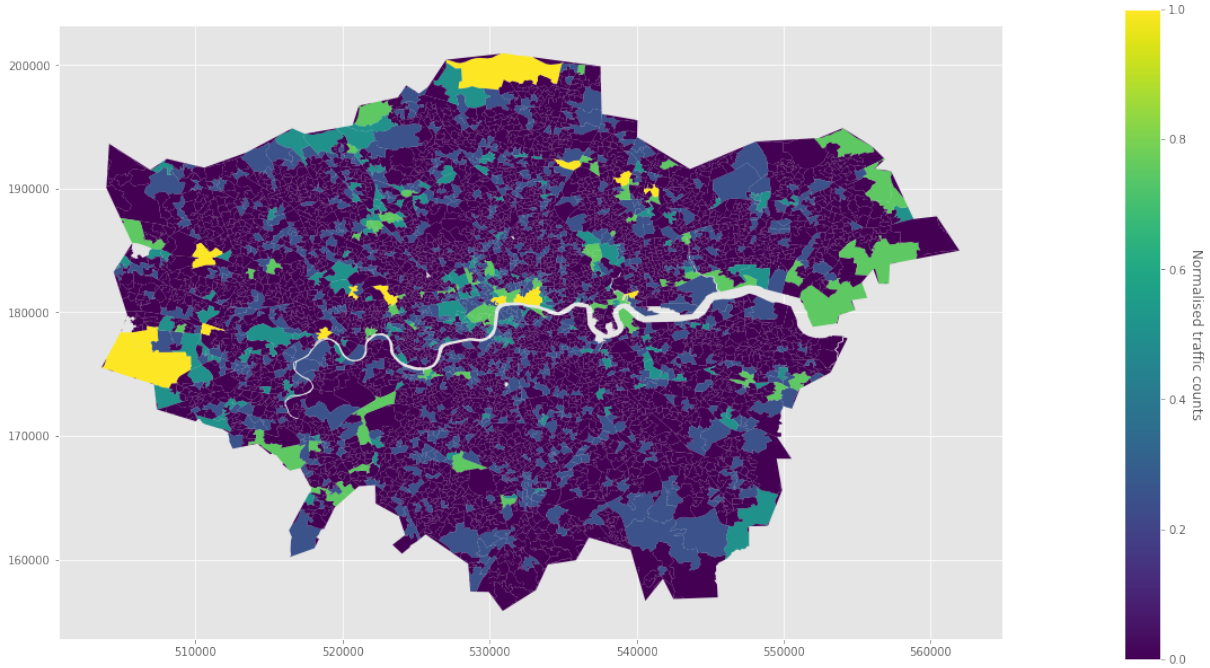


Figure 6.34: Standardised traffic counts

[4]The vehicle categories used were: Cars/Taxis, motorcycles, heavy goods vehicles, light good vehicles

6.4.2 Defining the capabilities

By definition, the set of capabilities should be constructed in a way that reflects an individual's choice to realise their desired goals, as well as the potential opportunities an individual has to make those choices. According to Hananel & Berechman (2016), an evaluation of the capabilities set should start by explicitly stating what these are. In the context of accessibility and adopting the definition of capabilities from Tyler (2006), these are framed around:

- the ability to engage with available opportunities and
- the ability to use the public transport system to do so.

The first bullet point is related to the probability distribution of activity types bounded by the isochrone polygon, while the second is related to the probability distribution of using the different public transport modes at each access point in a trajectory.

In particular, this case study will be investigating the following elements:

- Potential access to activities
- Potential mobility
- Public transport mode and activity type dynamics

6.4.2.1 Potential access to activities

This element of the capabilities set describes the potential range of activity types that are reachable within the walking time from a public transport access point defined by the activity space isochrone.

Similar to section 6.2, the effect of personal characteristics on the likelihood of reaching an activity type was captured through a Dirichlet distribution, using the concentration parameter vector α_z. This represents the degree of prior belief that an individual is likely to be performing one activity type over the other. For example, it might be that prior studies indicate that arrival time between 11:00-12:00 pm and age group < 21 years old can be used to determine education over employment activity. This assumption can be represented by setting $\alpha_{education} > \alpha_{employment}$.

Smaller ($0 < \alpha_z < 1$) values of α_z express less uncertainty in the preference of an activity type over the other. On the other hand, larger values ($\alpha_z > 1$) express more uncertainty about the preference of an individual for an activity type. In this study, the calculation of the shape of the prior was based on RODS data by

generating predicted probabilities for each activity type based on age, sex, disability status and arrival time. These were then used to construct the concentration parameter vector α_z. Similarly to section 6.3, the resulting predicted probabilities were multiplied with Gamma distributed random variables with shape and rate parameters of the Gamma distribution a = b = 1 to ensure that the concentration parameters follow an exponential distribution with rate proportional to the RODS predicted probabilities.

It has already been mentioned that, by definition, capabilities are not observed and should be treated as latent (hidden) quantities. Within quantitative studies in the CA literature and particularly within econometrics, different latent variable models have been applied to infer capabilities from observed functionings. Examples include dimensionality reduction methods such as Principal Component Analysis (PCA) and factor analysis as well as statistical models such as SEMs (eg. Generalised Linear Latent and Mixed (GLLM) models, MIMIC models) (Krishnakumar 2007, Anand et al. 2011). The advantages and disadvantages of these models have already been discussed in previous sections (section 4.6).

In the modelling specification of this case study, the set of potential activities is represented as a sequence of latent (unobserved) stochastic variables which, similarly to section 6.3, are inferred by:

- the combination of duration of stay and the number of defined POI types that are deemed reachable by foot from the public transport access point, as determined by the isochrone polygon and
- the propensity of performing each activity category based on personal characteristics as inferred from the RODS dataset.

The stochastic element allows the incorporation of uncertainty in the inference of activities as a function of the different configurations of the states of all other variables in the model.

6.4.2.2 Potential mobility

This element of capabilities describes the potential of public transport use from the modes that are available. Similar to potential accessibility, potential mobility is represented by a latent stochastic quantity that is inferred using:

- the propensity of public transport use given an individual's sociodemographic characteristics as inferred from the LTDS dataset and
- the observed public transport modes used throughout their trajectory from the Oyster card dataset.

In the context of this study, potential mobility was modelled using a categorical random variable over the Oyster card transportation mode types. This time, the propensity of an individual to use one mode over another was modelled using a multinomial regression on the LTDS dataset. In this case, the personal characteristics determining the choice of transportation mode were age, income, possession of travel pass, disability, car license, sex and ethnic group. The predicted probabilities were then recovered and used to shape the prior belief of using one mode over the others through the Dirichlet concentration parameters.

6.4.2.3 Public transport mode and activity type dynamics

As already mentioned in chapter 3, the notion of processes in the CA is central (Sen 1992). Processes can be best understood by considering the cumulative effects of time in people's beings or doings (Comim 2003). Within this framework, reaching destinations and interacting with public transport can be considered an evolving process that is shaped by an individual's choices over the available opportunities and transportation modes. This dynamic aspect has also been emphasised in accessibility studies, although approached from different angles (eg. through the evolution of the space-time prism in time geography).

A big challenge in both the CA and accessibility studies is the incorporation of this dynamic element in evaluative frameworks. Within the CA, this is due to a variety of reasons. First, the CA puts emphasis on the diversities among individuals in a society by looking at variations in personal characteristics, environmental conditions and personal resources. As Comim (2003) puts it, this focuses on interventions aimed at compensating or enabling the disadvantaged groups in some ways, which in turn focuses on comparisons using static snapshots of individual states of being or doing. Indeed, most studies that focus on change are limited to comparing relative numbers between different static states. Second, the undefined implementation framework of the CA makes it difficult for researchers to incorporate the dynamic element. This is of particular importance, since any dynamic evaluation needs to be made within the context of an individual's interaction with public transport/activities. Ignoring the evolution of an individual's personal characteristics/commodities and the environment could provide a false perception that a condition that leads to reduced accessibility is being addressed. Within accessibility studies, the dynamic element reflects changes in the transportation infrastructure as well as the attractiveness of destinations (Moya-Gómez et al. 2018) but could also reflect changes of an individual's circumstances such as time budget (Kwan 2013).

The underlying assumption that is made in this modelling step is that charac-

teristics of the environment have a varying effect on the ability of an individual to reach an activity. For example, the levels of deprivation change from location to location, and this is expected to influence the choice of performing an activity type at a particular location. Similarly, the existence of more transportation options (expressed as increased levels of transport accessibility) are expected to influence the choice of transportation modes.

For this case study, the approach followed to capture the transportation mode state dynamics in sections 6.2 and 6.3 was used to frame the dynamic nature of the proposed CAA. In particular, the transitions between the transportation modes (from n to $n-1$, where $n \in \{1...N\}$ and N is the sequence of Oyster card 'taps') were included through separate Multinomial regressions over those external factors. These separate row regressions were organised in a row stochastic transition matrix and used in the calculation of the likelihood of potential mobility. A similar approach was followed for modelling the transition dynamics of the different imputed activity types from the sequence of isochrone locations. This resulted in two square row stochastic matrices, a $5x5$ matrix for the 5 activity categories T_z and a $3x3$ matrix for the transportation modes T_m.

The posterior distributions of the Multinomial regression coefficients used to model the rows of the transition matrices can inform the predictive potential of the external factors on the output probabilities as an individual travels in space. In the context of this study, when referring to 'external factors', what is meant is the external covariates within the regression framework. Within the CA literature, the use of a regression framework to estimate the degree of contribution of different external factors is not new. For example, Krishnakumar (2007) used the estimated coefficients of a SEM to elaborate on the exogenous factors influencing human development variables such as health, knowledge and political freedom. Similarly, Di Tommaso (2007) used the estimated coefficients of a MIMIC model to assess the influence of external factors on capabilities such as health and leisure activities.

For this modelling specification, the output probabilities of the transition matrices correspond to the transition potential between the different PT modes and the different activities. In a Bayesian setting, a normal prior distribution is placed on each coefficient for each row category:

$$\beta_i^{\kappa} \sim Normal(0, 10^{-3}) \tag{6.11}$$

for $\kappa = \{1...K\}$ categories and $i = \{1...\#covariates\}$ external variables. The external variables included in modelling of transportation modes was *Trip duration, PTAL*, while for activities the external variables included were *Crime, Green areas, Traffic, IMD.*

The use of transition matrices to capture the dynamic component of a system is not new within the wider urban systems literature and has been used in different contexts such as microsimulation of sociodemographic effects of individuals in transport demand models (Goulias & Kitamura 1992), simulation of the future urban land use scenarios by including the density of transport in the calculation of transition matrix (Barredo et al. 2003) as well as modelling the co-evolution of land use and transport provision (Levinson & Chen 2005). Regarding the coefficients of explanatory variables to inform on the effect of different external factors on capabilities, there are numerous examples within the CA literature supporting this approach (Hickman et al. 2017, Anand et al. 2011, Krishnakumar 2007).

6.4.3 Defining the functionings

The above described capabilities are mapped to the attainable functioning vectors a person can achieve given personal characteristics/external variables. These are measurable quantities which are observed and used to infer the unknown quantities of the latent capabilities variables.

For this case study, functionings are considered to be an individual's realised mobility behaviour as evidenced by the Oyster/LTDS card sequence of "taps", which are translated to the public transport modes an individual can use at a particular point in time. The available opportunities on the other hand, are translated to all potential activity types an individual can reach from a particular public transport access point by walking. The two functioning vectors are then:

$$
\begin{aligned}
m^{\kappa} &= \{Bus,\ Rail\ Tram\} \\
z^{\kappa} &= \{Employment,\ Sports\ and\ Entertainment, \\
&\quad Education\ and\ Health,\ Eating\ and\ Drinking,\ Retail\}
\end{aligned}
\tag{6.12}
$$

z^{κ} is an observed stochastic variable assumed to follow a multinomial distribution over the isochrone defined POI vector counts. The transportation mode vector m^{κ} is assumed to follow a categorical distribution with the probability parameter p being determined by the transition matrix T_m and the effect of personal characteristics.

6.4.4 Bringing it all together: Defining the structure of the CAA model using Bayesian networks

The CAA model consists of two distinct but intertwined modules: 1) activity inference and modelling and 2) mobility modelling. These are combined using

a switch variable that activates the relevant module depending on whether an individual is using public transport, or assumed to be performing an activity. Figure 6.35 illustrates a graphical representation of the joint model.

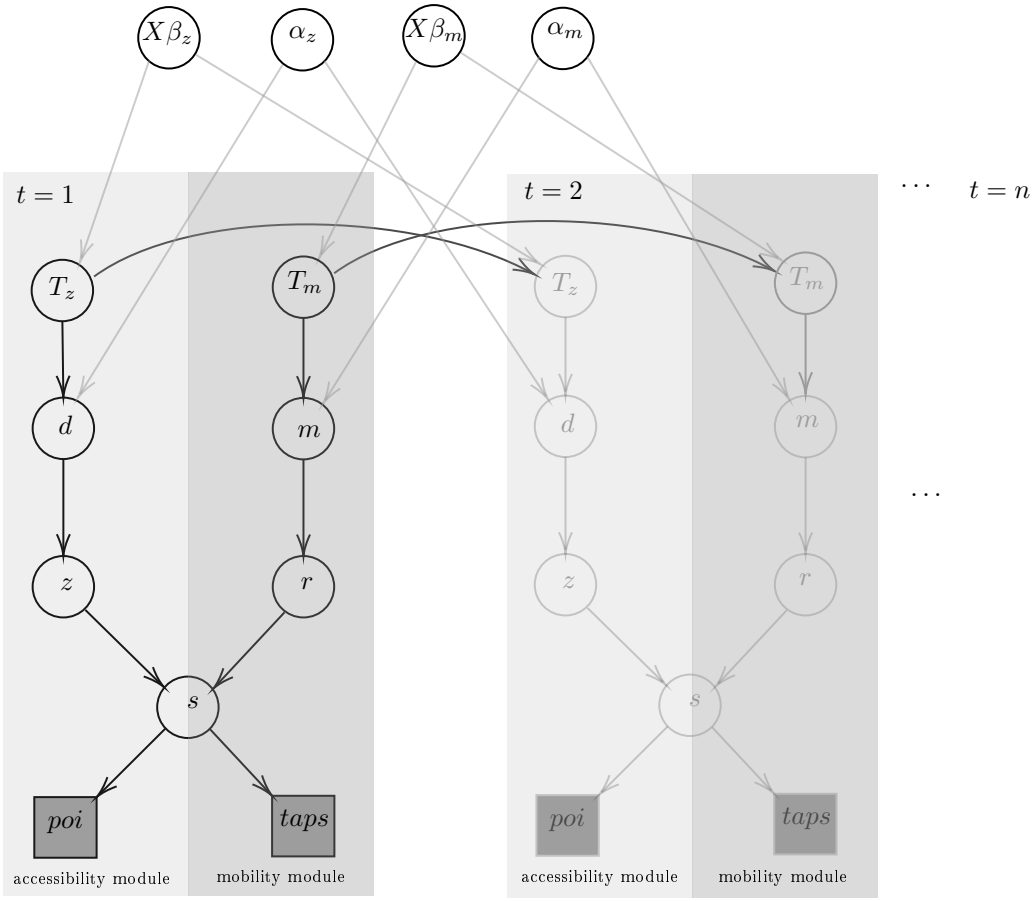


Figure 6.35: Graphical representation of the dynamic Bayesian network used for activity inference.

Similarly to previous chapters, the individual's personal characteristics are included as stochastic nodes on the top level hierarchy of the dynamic Bayesian network for both inference modules. This reflects the idea that information encoded in these variables propagates to the subsequent nodes of the network. Within the context of the network, these are stochastic variables that follow a prior distribution that is shaped from personal characteristics. Similarly to section 6.3, the prediction results from the fitted LTDS/RODS driven regressions using an individual's personal characteristics were multiplied with Gamma distributed random variables ($a = b = 1$). For the prediction probabilities, the softmax function was used:

$$s^{1...\kappa} = \frac{e^{X'\hat{\beta}}}{\sum_{\kappa=1}^{K} e^{X'\hat{\beta}_\kappa}} \tag{6.13}$$
$$\alpha^{1...\kappa} \sim s * Gamma(a, b)$$
$$p \sim Dirichlet(\alpha^1 ... \alpha^\kappa)$$

where c are the predicted probabilities. $\hat{\beta}$ are the estimated LTDS/RODS regression coefficients.

External variables are included in the transition matrices T_z and T_m for activity type and transportation mode modelling respectively. For T_z, this included Crime, Proportion of Green areas, Traffic and IMD. For T_m this included travel duration and PTAL. The figure below shows the regression submodel used to construct the rows of the transition matrices:

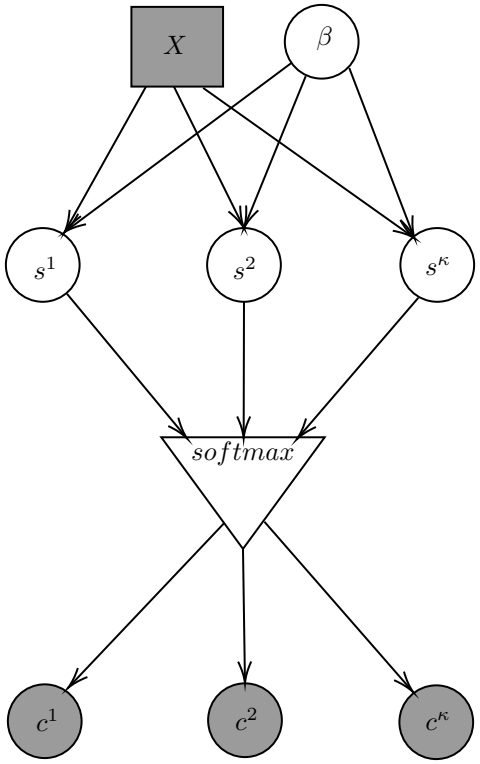


Figure 6.36: Row transition estimation using external factors

where X is the external variable design matrix and $softmax$ represents the softmax function.

The transition sequences for the inferred activities/transportation modes specified per each category were constructed as follows:
where $y_i = argmax(z_i)$ in the case of activities, and $y_i = m_i$ in the case of transportation mode.

Algorithm 1 Construction of transition sequence

```
procedure CONSTRUCT TRANSITION SEQUENCE(input = c^{1...κ})
    for i in 1 : N do
        for κ in K do
            if y_i = κ then
                append y_{i-1} to c^κ
```

Essentially, the algorithm generates a transition dataset from one category to another by looping through the trajectory locations and identifying if a transition between location n and location $n-1$ is related to activity types/transportation mode k. For example, consider a trajectory with transportation modes bus_1, bus_2, $rail_3, bus_4$. In this case, the row of the transition matrix corresponding to bus related transitions will be inferred using the sequence $bus, rail$ as there is one bus/bus related transition (from $n = 1$ to $n = 2$) and one $bus/rail$ related transition (from $n = 2$ to $n = 3$)

Variables $c^1, c^2 ... c^\kappa$ were then assumed to follow a Categorical distribution over the above transition sequences:

$$\begin{aligned} \beta &\sim Normal(0, 10^{-3}) \\ s^{1...\kappa} &= \frac{exp(\alpha + \beta X_n)}{1 + \sum_{\kappa=1}^{K-1} exp(\alpha + \beta_\kappa X_n)} \\ c^{1...\kappa} &\sim Categorical(s^{1...\kappa}) \end{aligned} \tag{6.14}$$

The transition matrices T_m and T_z are then constructed using the inferred s^κ:

$$T = \begin{bmatrix} s^{\kappa=1} \\ s^{\kappa=2} \\ \vdots \\ s^{\kappa=K} \end{bmatrix} \tag{6.15}$$

Furthermore, the activity related personal characteristics α_z were included directly in the computation of the likelihood of d, as it is assumed to directly influence the probability of activity within an isochrone:

Algorithm 2 Computation of the likelihood of T_m, T_z, m and d

1: **procedure** COMPUTE T_m LIKELIHOOD($input = s_m^{1...\kappa}$)
2: logp = 0
3: **for** row in T_m **do**
4: logp =+ dirichlet likelihood($value_{row}$, $s_m^{1...\kappa}$)
5: **procedure** COMPUTE T_z LIKELIHOOD($input = s_z^{1...\kappa}$)
6: logp = 0
7: **for** row in T_z **do**
8: logp =+ dirichlet likelihood($value_{row}$, $s_z^{1...\kappa}$)
9: **procedure** COMPUTE d LIKELIHOOD($input = T_z, \alpha_z$)
10: logp = dirichlet likelihood(value, α_z)
11: P = Unconditional Probability(T_z)
12: logp =+ dirichlet likelihood($value_0$, P)
13: **for** i *in* $1:\ N$ **do**
14: logp =+ dirichlet likelihood($value_i$, $T_{row=argmax(d_{i-1})}$)
15: **procedure** COMPUTE m LIKELIHOOD($input = T_m, \alpha_m$)
16: logp = dirichlet likelihood(value, α_m)
17: P = Unconditional Probability(T_m)
18: logp =+ dirichlet likelihood($value_0$, P)
19: **for** i *in* $1:\ N$ **do**
20: logp =+ dirichlet likelihood($value_i$, $T_{row=argmax(m_{i-1})}$)

In the context of MCMC inference with Metropolis-Hastings algorithm, the *value* parameter in Algorithm 2 is generated by a normal proposal distribution (or Poisson in the case of discrete variables) and acceptance/rejection is evaluated according to the computed likelihood. The function *Unconditional Probability* was computed using equation 6.1.

Moving down the causal structure of the model, the latent activities are represented by:

$$\begin{aligned} P(d_n|\alpha_z, T_z) &\sim Dir(\alpha_z, T_{row=argmax(d_{n-1})}) \\ P(z_n|d_n) &\sim Mult(poi_n, d_n) \end{aligned} \tag{6.16}$$

where $Dirichlet(\alpha)$ represents a Dirichlet distribution defined by a concentration parameter vector α and $Mult$ represents a multinomial distribution with probability vector d and a total count of POI elements n.

Since, given the nature of the dataset, there are no observations on the choice of activity by an individual, these are imputed from the duration of stay and the POI counts inside an isochrone area. The assumption is then that $P(z^{\kappa}|d)$ defines the functionings, representing what can potentially be achievable inside the isochrone area given the POI/duration dataset. Algorithm 2 provides the pseudo-code for computing the likelihood of d node.

Similarly to chapter 6.2 the extent of personal influence is determined by the predicted probabilities of performing an activity given the personal characteristics as determined by the RODS dataset (e.q. 6.13).

An individual's mobility (in this case study the use of public transport), is represented by a categorical distribution $r \sim Cat(m)$ over the sequence of Oyster card "taps". Similarly to z, the probability vector p over each "tap" was modelled through a latent Dirichlet variable m. Doing so, allows the modelling the different transportation modes through a probability distribution for each public transport access point. The effect of personal characteristics α_m was also included in the likelihood computation as, similarly to d, they are assumed to persist throughout an individual's trajectory:

$$\begin{aligned} P(m_n|\alpha_m, T_m) &\sim Dir(\alpha_m, T_{row=argmax(m_{n-1})}) \\ P(r_n|m_n) &\sim Cat(m_n) \end{aligned} \tag{6.17}$$

In this case, the notion of functionings is more straightforward as the Oyster card "taps" are direct observations of the actual choice of public transport mode made by the individual.

Finally, the node b is a stochastic variable acting as a switch that controls which module is activated for inference (accessibility or mobility). It is assumed to follow a Bernoulli distribution, the probability of which is determined by the duration of stay relative to the cutoff value determined from the 95th percentile of the distribution of interchange times for bus and rail services (in the case of rail services, this was 15 minutes while for buses this was 36 minutes). For example, if the duration of stay between two subsequent bus trips is more than 36 minutes, then it is more likely that an activity is carried out at the stop (as opposed to being an interchange stop).

$$P(b) \sim Bernoulli(p) \tag{6.18}$$

At the very bottom of the hierarchy of Figure 6.35 the square nodes represent the observed mobility and POI data used to infer the parent nodes.

The MCMC sampling step methods used for each node of the model were:

- $b, z, r, \beta_m, \beta_z, c^{1\ldots\kappa}, s^{1\ldots\kappa}, \beta, \alpha_m, \alpha_z$: Metropolis-Hastings
- d, m, T_z, T_m: Adaptive Metropolis

A total of 20000 sampling iterations were made (2 runs of 10000 samples each), with starting values of the stochastic variables sampled from the prior distributions. The simulations were stored in separate SQLlite databases with size ≈500MB for each individual, while the complete simulation required ≈100 minutes to complete per individual trajectory. Convergence was achieved for the majority of nodes in the model, however, nodes that are characterised by deterministic components and increased collinearity were more difficult to sample. In particular, sampling from the transition matrices was more challenging, most probably due to the complicated likelihood of this node. To overcome this problem, a combination of different sampling schemes was used for each node in the model together with adjusting the scale of MCMC algorithm depending on the acceptance rate. Furthermore, an accuracy assessment was performed on the activity type detection component of the CAA model with 9 participants. In general, the performance of the model was comparable to the results of section 6.3, this time however, higher overall accuracy could be achieved due to increased predictability of *Employment* and *Education* activities. Detailed convergence diagnostics are given in Appendices D and E.

6.5 Chapter summary

In this chapter, the technical implementation of the CAA was introduced in section 6.4 building on the model specifications of sections 6.2 and 6.3. This model was applied using unlabelled, passive mobility data from London's AFC system together with travel survey data. The elements of the capabilities set were explicitly defined and linked with an individual's personal characteristics, external variables and functionings using a dynamic Bayesian network structure.

The following chapter (chapter 7) presents the results for all nodes of the CAA model using the Oyster/LTDS data. These are then used within the context of evaluating the risk of transport social exclusion experienced by individuals when using the public transport to reach their day to day activities.

Chapter 7

Assessing transport related social exclusion using a capabilities approach to accessibility model: Results

7.1 Chapter overview

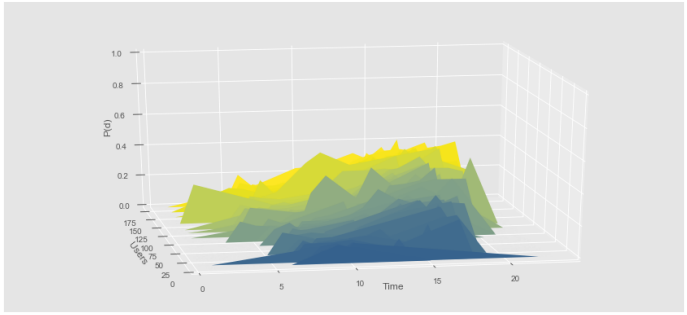
In this chapter, the individual accessibility patterns resulting from the output of the modelling framework of chapter 6 are discussed, particularly in relation to individuals belonging to the unconstrained (baseline) group. The results are then used to formulate a framework for assessing relative transport disadvantage and social exclusion experienced by individuals belonging to the unconstrained, >60 and low income population group (section 6.4). For this task, a popular entropy based equality index is applied to the posterior quantities of the Capabilities Approach to Accessibility (CAA) model. Finally a discussion of the results in relation to the distribution of the individual equality indices is offered, aiming to highlight aspects of social exclusion and transport disadvantage experienced by some individuals. The analysis and discussion of this chapter has been published in the article "Assessing transport related social exclusion using a Capabilities Approach to accessibility framework: A dynamic Bayesian network approach" of journal Journal of Transport Geography (Bantis and Haworth 2020).

7.2 A Capabilities approach to accessibility model: Results

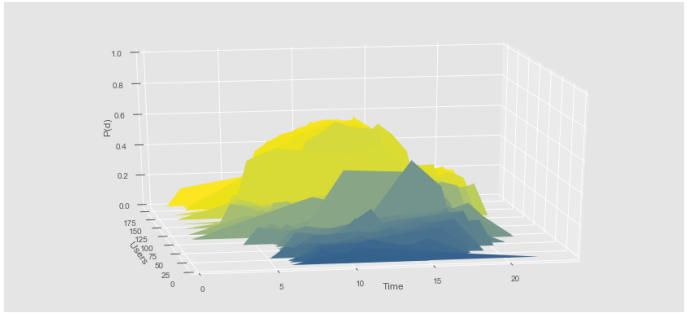
This section describes the results for all the core variables of the CAA model. At the end of each subsection, a more consolidated description of the results is given, by enumerating the main findings for each variable.

7.2.1 Distributions of activity types

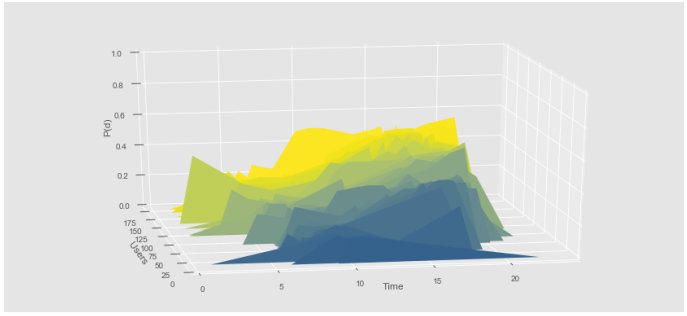
The use of passive trajectory data in the CAA modelling framework enables the evaluation of activity type probabilities for each individual in the sample throughout the day. The posterior distribution of activity types corresponds to the latent d node, which expresses the posterior distributions of activity types given the individual's sociodemographic characteristics, duration of stay and number of reachable POIs from the alighting point. Figure 7.1 below shows the posterior quantities of $P(d)$ for the trajectories of all users in the unconstrained Oyster card sample. The *x-axis* (labelled *Time* in the figure) shows the time of day throughout an individual's trajectory while the *z-axis* (labelled $P(d)$ in the figure) shows the probability of performing an activity type throughout the day. The *y-axis* (labelled *Users* in the figure) shows the corresponding distribution for each individual in the sample.



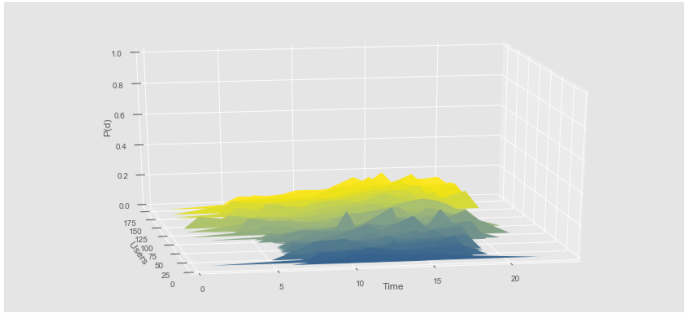
(a) Eating and drinking



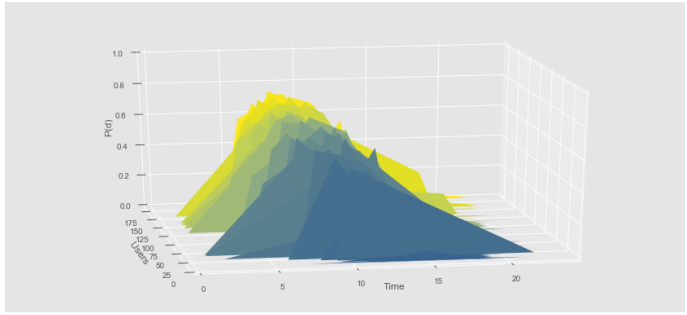
(b) Education and health



(c) Retail



(d) Outdoors and Recreation



(e) Employment

Figure 7.1: Posterior distributions of activity types for the unconstrained population sample

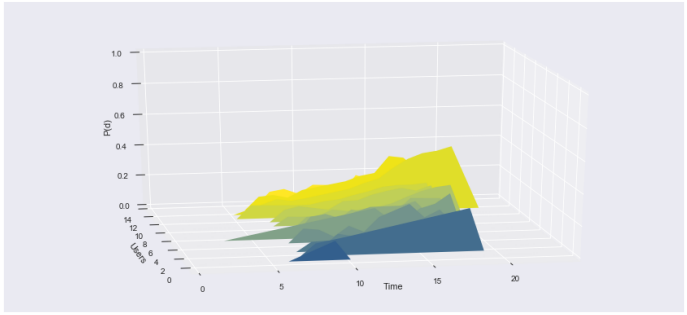
It can be seen that there is a similarity between *Eating and Drinking* (figure 7.1a) and *Retail* (figure 7.1c) activity types, reflecting the overlap between the POIs belonging to these categories inside an activity space isochrone. For both of these activity types, the posterior probabilities increase throughout the day for the majority of individuals, peaking around mid-afternoon. This reflects the intuition that these types of activities tend to be performed later in the day. For the posterior distributions of activity type *Education and Health* (figure 7.1b) there seems to be a within-category split throughout the day, with the probability for some individuals performing this activity type peaking in the morning, while for others it peaks later in the afternoon or persists throughout the day. An explanation for this pattern could be the aggregated nature of this activity type, as education activities tend to peak in the morning while health activities tend to persist throughout the day.

The posterior probabilities for *Employment* (figure 7.1e) are considerably higher in the time window between 8:00-10:00AM compared to the rest of the activity types, reflecting the start of an individual's working day. The probabilities then gradually decrease throughout the day, mirroring the increasing pattern of the rest of the activity types.

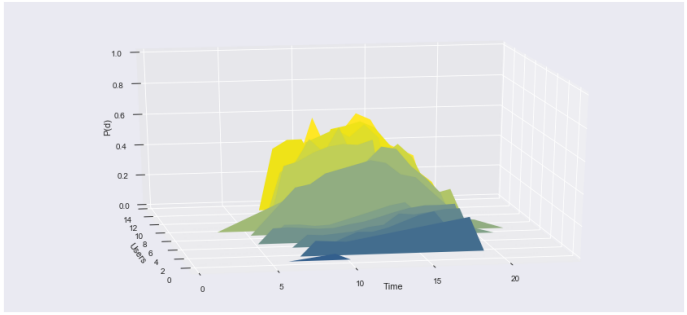
On the other hand, the posterior probabilities for activity type *Outdoor and Recreation* (figure 7.1e) almost never reach a level above the random probability assignment, given the total number of activity categories ($P(d) < 0.2$). An possible interpretation of this result is the lack of POI data and the infrequency of performing this activity compared to the rest of the POI categories, particularly considering the use of public transport.

The posterior quantities for the individuals in the low income group and the $>$60 years old group were similar to the unconstrained group for all categories (figures 7.3,7.2). For the low income group one notable difference is the shorter tails of the distributions for the majority of the individuals for the *Employment* activity (figure 7.2e). This could signify reduced flexibility in using public transport to reach this activity compared to the unconstrained population group. Moreover, the probabilities of *Eating and Drinking* (figure 7.2a) and *Retail* (figure 7.2c) activity types is significantly lower throughout the day, remaining below the threshold for random probability allocation for the specified number of activity types (< 0.2). Contrary to the rest of the population groups, for the > 60 years old group the *Education and Health* (figure 7.3b) category is characterised by a gradual increase over the later hours of the day for the majority of the individuals. This could be attributed to health related activities as opposed to *Education* related activities. This is in contrast to the low income group where *Education and Health* activ-

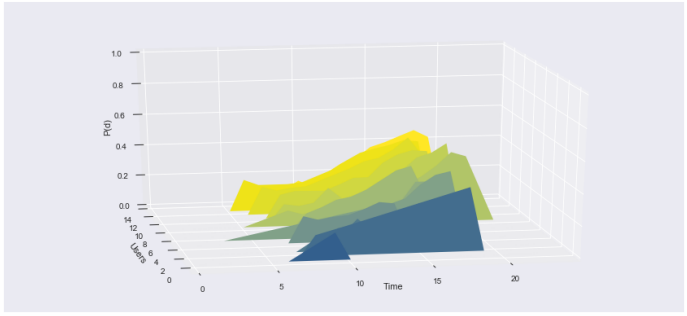
ity for a significant number of individuals is peaking between 10:00AM-15:00PM (figure 7.2b). Together with *Employment* activity, the posterior probabilities for *Education and Health* dominate the rest of activity types, signifying the presence of students in the low income population group. The *Employment* (figure 7.3e) activity type for the > 60 years old group seems to dominate the daily trajectory of this group for the early hours of the day. This is not surprising considering the fact that the majority of the individuals in this group were below the UK national pension age (63 years for women and 65 years for men). Nevertheless, a general shift of this activity type can be observed to slightly later hours of the day compared to the rest of the target groups of this study, reflecting some flexibility in using the public transport to access employment.



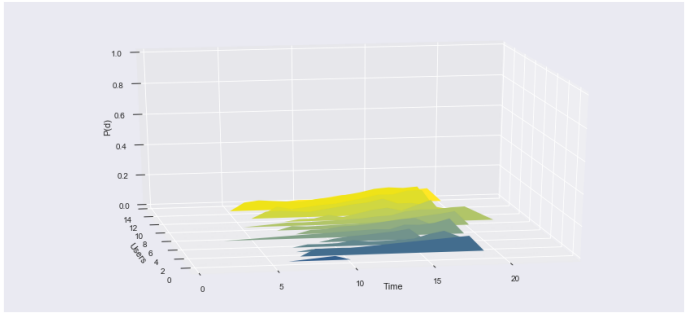
(a) Eating and drinking



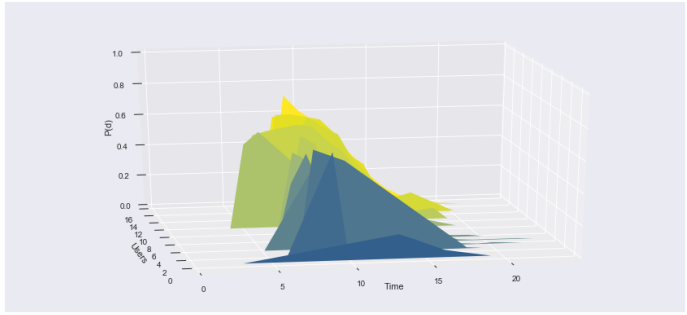
(b) Education and health



(c) Retail

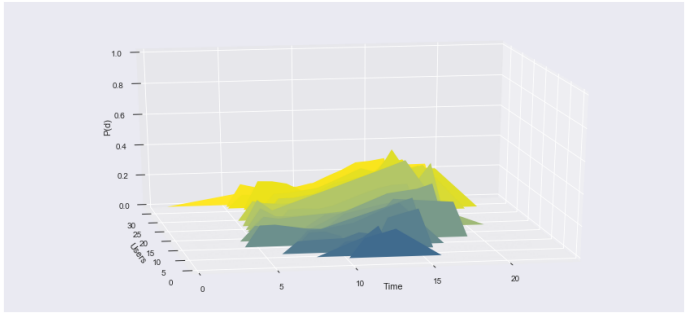


(d) Outdoors and Recreation

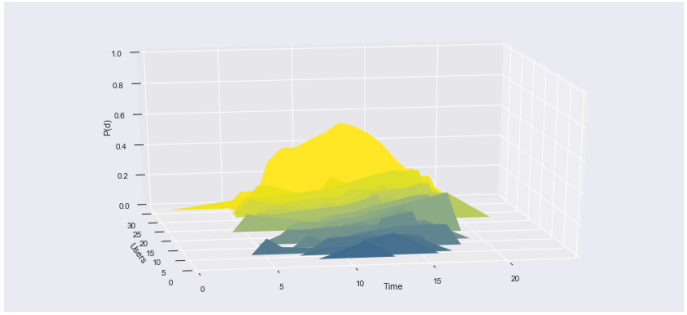


(e) Employment

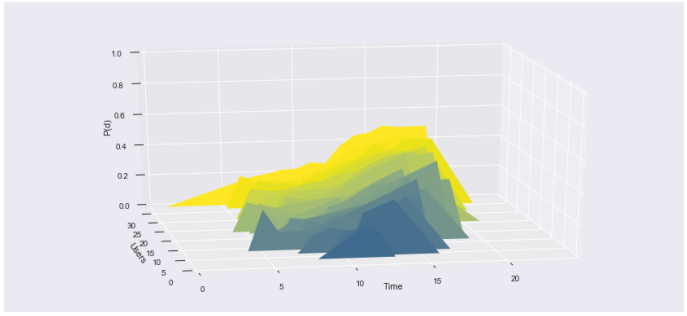
Figure 7.2: Posterior distributions of activity types for the low income population sample



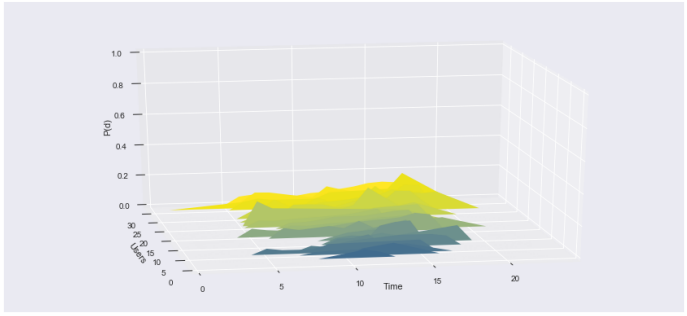
(a) Eating and drinking



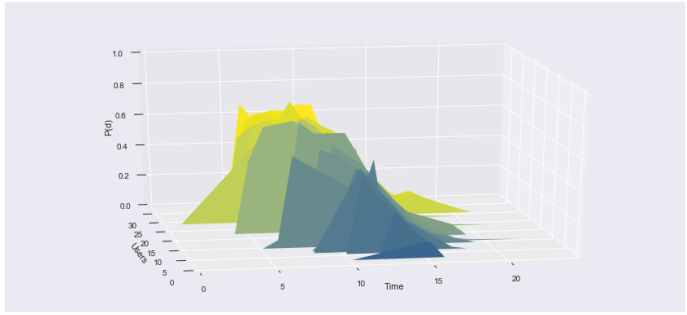
(b) Education and health



(c) Retail



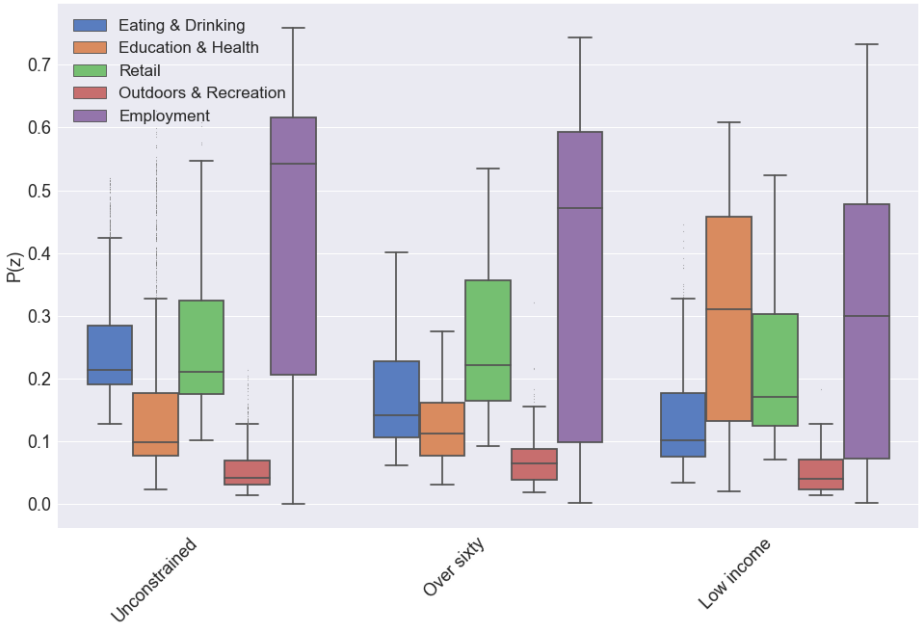
(d) Outdoors and Recreation



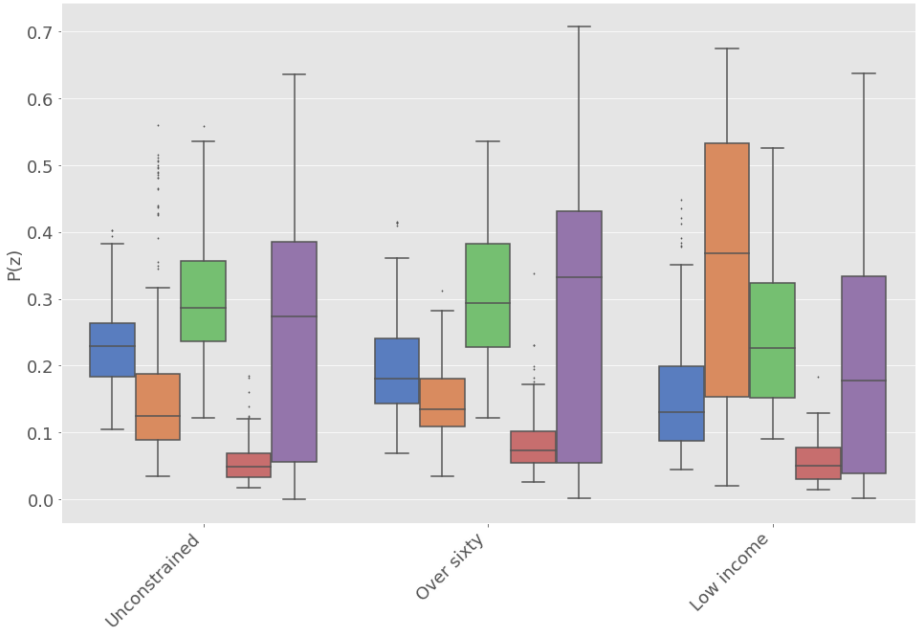
(e) Employment

Figure 7.3: Posterior distributions of activity types for the over sixty years old population sample

Figure 7.4 shows aggregated boxplots of posterior distributions for the different activity types, for all individuals in the target groups for both weekdays and weekends. As it can be seen, the general pattern of activity distribution remains with the exception of employment activity which is considerably lower on the weekends.



(a) Aggregated activity distribution boxplots for all individuals (weekdays)



(b) Aggregated activity distribution boxplots for all individuals (weekends)

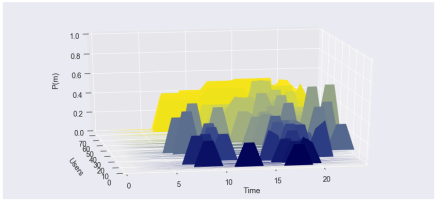
Figure 7.4: Aggregated activity type boxplots for the three population groups.

Summarising, the findings for this posterior quantity are:

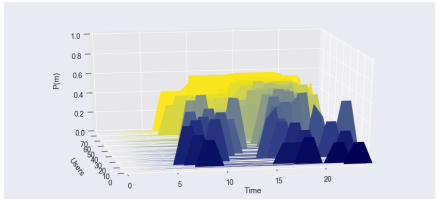
- Non-employment probabilities for Eating and Drinking and Retail activities increase in the later hours of the day for the unconstrained group.
- Employment probabilities are higher in the morning hours for the unconstrained group.
- Low income group employment distributions are narrower compared to the unconstrained group.
- Low income non-employment activity probabilities are lower compared to the rest of the population groups.
- > 60 group education and health probabilities increase in the later hours of the day.
- > 60 group employment activity appears to be more flexible throughout the day.

7.2.2 Distributions of transportation modes

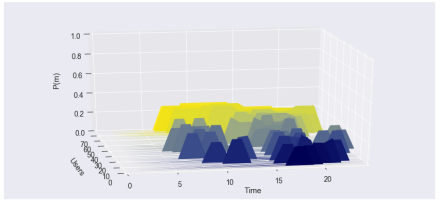
The next posterior quantity of interest is the distribution of transportation modes for each individual in the focus population groups. Similarly to section 7.2.1, the use of individual trajectories enables the evaluation of transportation modes used by individuals throughout the day. This corresponds to the latent m node of model 6.35 and relates to the mobility element of the defined capabilities. Similarly to the d node, the results for the unconstrained population group are presented per activity type (figure 7.5). The segmentation per activity type was made by taking the activity with the highest probability from each individual activity distribution at a visited location.



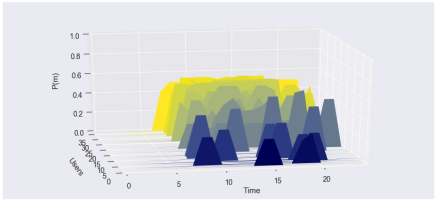
(a) Posterior means Bus/Eating and Drinking



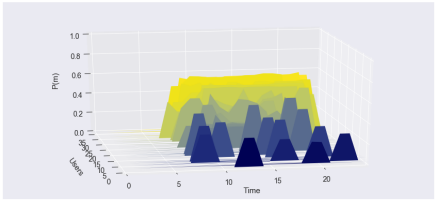
(b) Posterior means Rail/Eating and Drinking



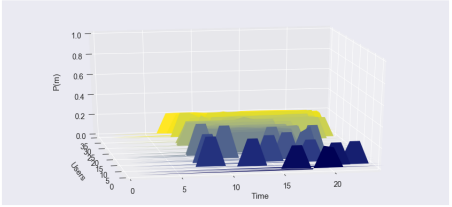
(c) Posterior means Tram/Eating and Drinking



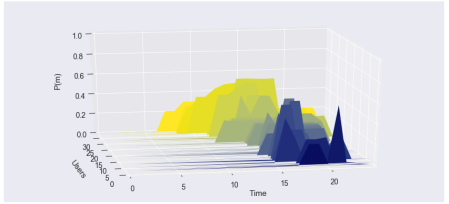
(d) Posterior means Bus/Education and Health



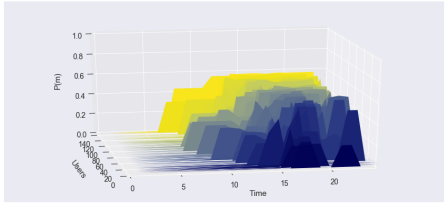
(e) Posterior means Rail/Education and Health



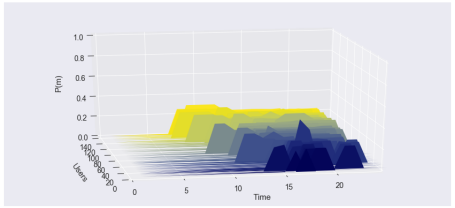
(f) Posterior means Tram/Education and Health



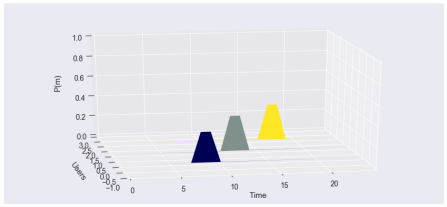
(g) Posterior means Bus/Retail



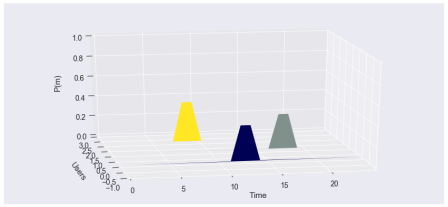
(h) Posterior means Rail/Retail



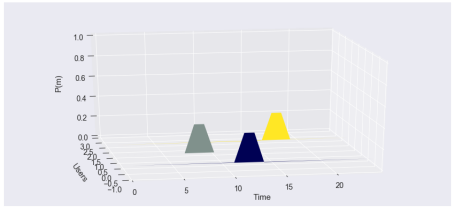
(i) Posterior means Tram/Retail



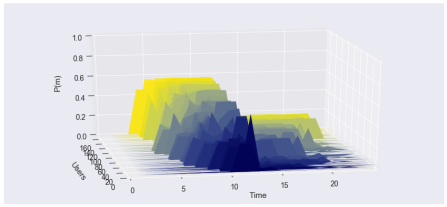
(j) Posterior means Bus/Outdoors and Recreation



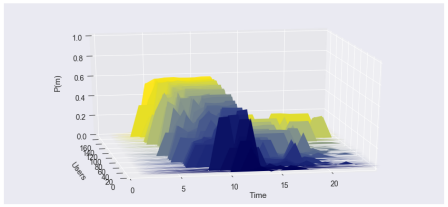
(k) Posterior means Rail/Outdoors and Recreation



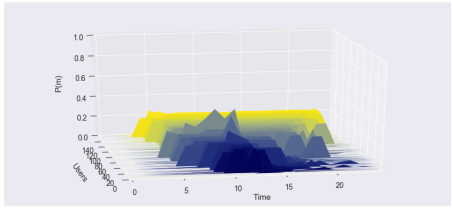
(l) Posterior means Tram/Outdoors and Recreation



(m) Posterior means Bus/Employment



(n) Posterior means Rail/Employment



(o) Posterior means Tram/Employment

Figure 7.5: Posterior means for the unconstrained population group

Looking at activity type *Eating and Drinking* (figures 7.5a, 7.5b, 7.5c), individuals in the unconstrained population group appear to be using the bus services uniformly throughout the day, with the probabilities increasing in the afternoon/evening. Individuals accessing this activity using rail services appear to form two distinct temporal clusters, the first one characterised by high probabilities in the late morning hours with the second by peaking in the afternoon/evening, a pattern which could be associated with lunchtime/leisure hours. Posterior probabilities for the tram service appear to be relatively low compared to the rest of transportation modes. This can be attributed to the absence of tram observations in the sample, reflecting the limited spatial coverage of this service. This is further discussed in section 8.4 of this thesis.

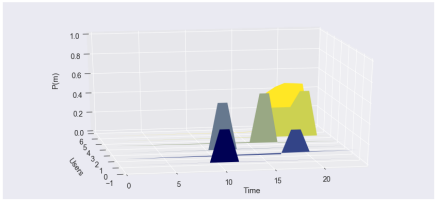
The situation with travelling by bus is similar for activity *Education and Health* (figures 7.5d, 7.5e, 7.5f). However, here the temporal clustering observed for rail services in the *Eating and Drinking* activity type is missing. Moreover, there appears to be a grouping between individuals characterised by relatively low probabilities for approximately half of them ($0.45 > P(m) > 0.25$) and relatively high probabilities for the other half ($P(m) > 0.45$). In addition, fewer people in this population group were found to be performing this activity type (35 individuals in total. The exception is the *Outdoors and Recreation* activity type).

A very similar pattern between bus and rail services is observed with activity type *Retail* (figures 7.5g, 7.5h, 7.5i). Here, the probabilities gradually increase throughout the day, peaking in the afternoon/evening hours for the majority of individuals, following the general pattern of shop opening times. For this activity type, rail seems to be the dominant transport mode.

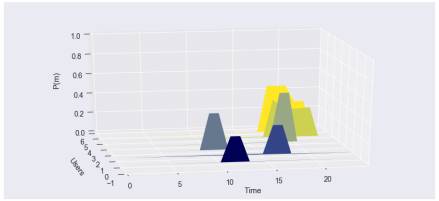
Finally, figures 7.5m 7.5n and 7.5o show the posterior results for activity type *Employment*. It can be seen that both bus and rail services are characterised by a similar pattern, defined by the morning commuting to work activity which, compared with the rest of the activity types, occupies a narrower temporal window of high probabilities between 7:00-11:00AM, levelling off in the afternoon/evening. Compared to using the bus, there is a slightly higher probability of using rail to reach this activity for the majority of individuals in the sample. Similar to the rest of the activity types, for tram services the probabilities are significantly lower, reflecting the lack of tram related transactions in the Oyster card sample.

For activity type *Outdoors and Recreation* the posterior probabilities of using the different transportation modes remained random ($P(m) \approx 0.33$). This can be attributed to the low posterior probabilities for the model's d node.

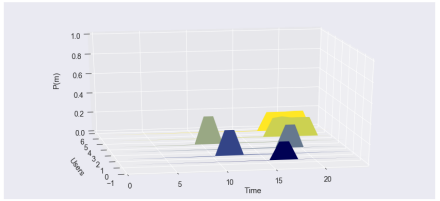
Next the transportation mode posterior distributions for the low income group is shown in figure 7.6.



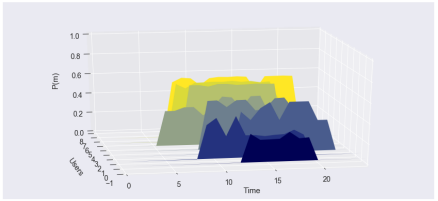
(a) Posterior means Bus/Eating and Drinking



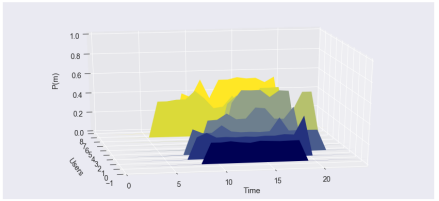
(b) Posterior means Rail/Eating and Drinking



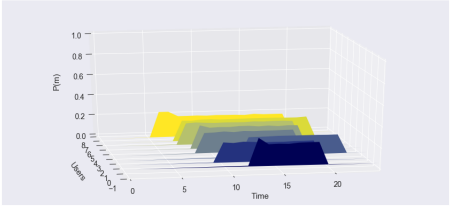
(c) Posterior means Tram/Eating and Drinking



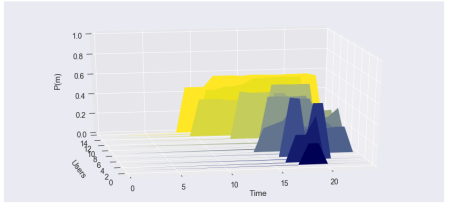
(d) Posterior means Bus/Education and Health



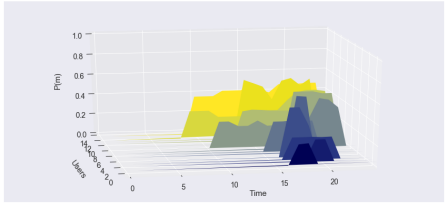
(e) Posterior means Rail/Education and Health



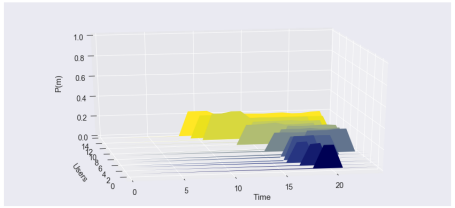
(f) Posterior means Tram/Education and Health



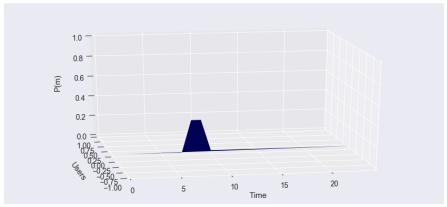
(g) Posterior means Bus/Retail



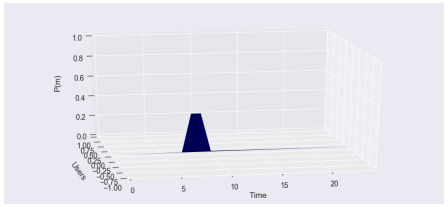
(h) Posterior means Rail/Retail



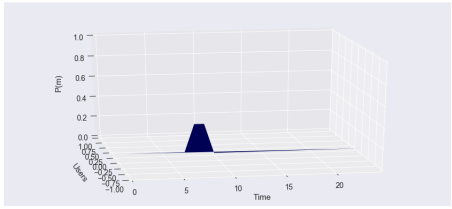
(i) Posterior means Tram/Retail



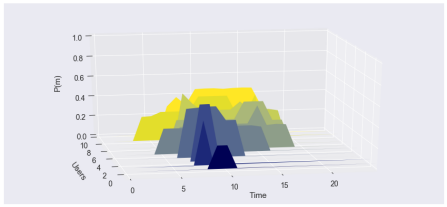
(j) Posterior means Bus/Outdoors and Recreation



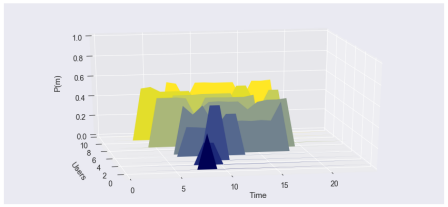
(k) Posterior means Rail/Outdoors and Recreation



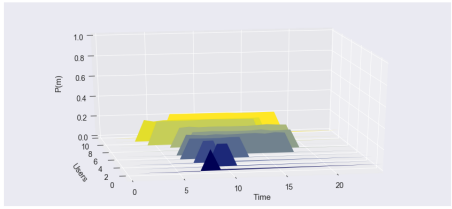
(l) Posterior means Tram/Outdoors and Recreation



(m) Posterior means Bus/Employment



(n) Posterior means Rail/Employment

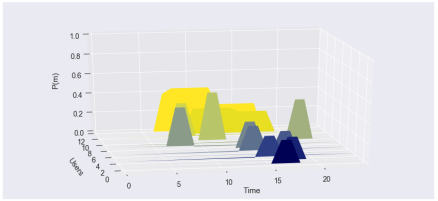


(o) Posterior means Tram/Employment

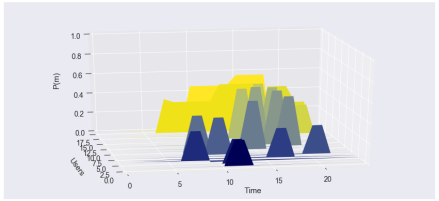
Figure 7.6: Posterior means for the low income population group

Compared to the unconstrained population group, the probabilities of using public transport to reach activity type *Eating and Drinking* are significantly lower, with nearly half of individuals in this group assigned to this activity type. Travelling by bus is the predominant transport mode for this activity type (figures 7.6a, 7.6b,7.6c). The posterior probabilities for *Education and Health* activity type (figures 7.6d,7.6e,7.6f) are slightly higher for using the bus compared to rail services, and the same holds for the *Retail* activity type (figures 7.6g,7.6h,7.6i). In terms of the overall shape of the distributions, *Retail* seems to follow the trend observed in the unconstrained population group, coinciding with retail shops' most popular times. Looking at the *Employment* activity type (figures 7.6m,7.6n,7.6o), the shape of posterior distributions for bus, rail and tram services is significantly wider than the unconstrained sample, characterised by two peaks. These are in the morning and early afternoon, most likely reflecting the varying schedule of part-time workers. Finally, for this population group, only one individual was attributed with reaching the *Outdoors and Recreation* activity type with higher probability of using the rail services.

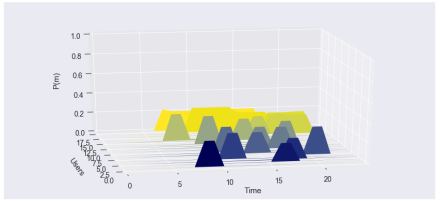
For the > 60 population group (figure 7.7), the probability of using the bus and using rail to reach activity type *Eating and Drinking* seem to compliment each other, with the probabilities of bus use being higher in the morning/afternoon and rail being higher in the afternoon/evening hours (Figures 7.7a, 7.7b). Overall, the probabilities of using the bus versus rail is similar for *Education and Health*, with rail services appearing to have a slightly shifted distribution mode toward the afternoon hours (figures 7.7d, 7.7e). The pattern of use of different transportation modes to reach *Retail* (figures 7.7g, 7.7h, 7.7i) appears to be similar to the rest of population groups. However, here the distribution of transport modes used throughout the day appears to be wider for a significant number of individuals. Moreover, compared to the rest of the groups of interest, more people are found to be using the bus for reaching *Outdoors and Recreation* activities.



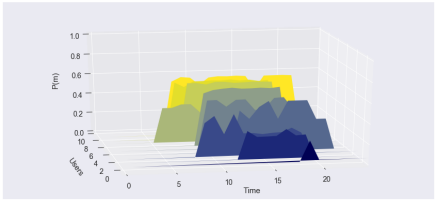
(a) Posterior means Bus/Eating and Drinking



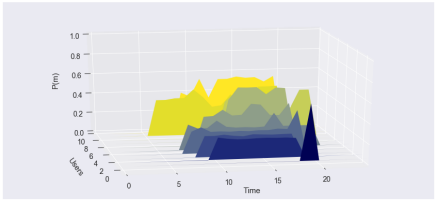
(b) Posterior means Rail/Eating and Drinking



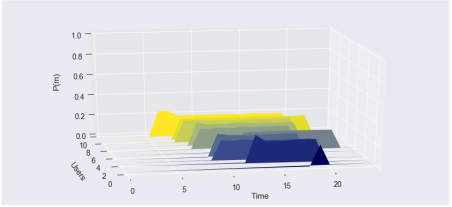
(c) Posterior means Tram/Eating and Drinking



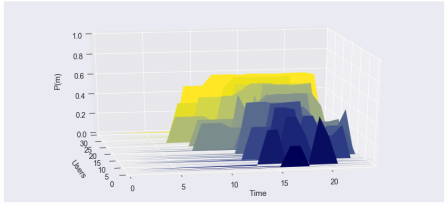
(d) Posterior means Bus/Education and Health



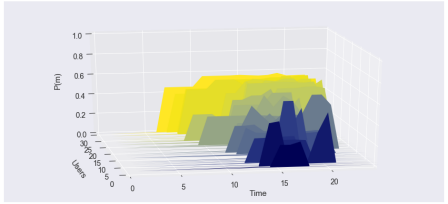
(e) Posterior means Rail/Education and Health



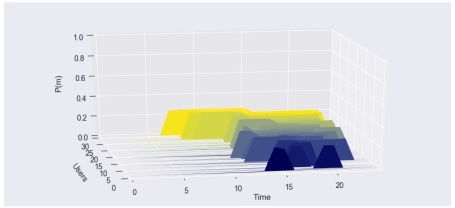
(f) Posterior means Tram/Education and Health



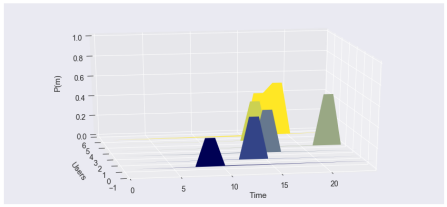
(g) Posterior means Bus/Retail



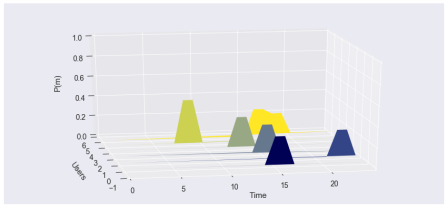
(h) Posterior means Rail/Retail



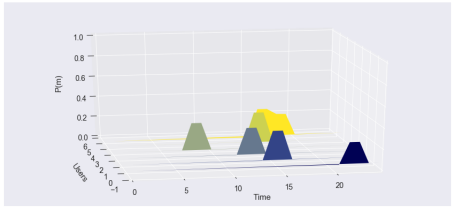
(i) Posterior means Tram/Retail



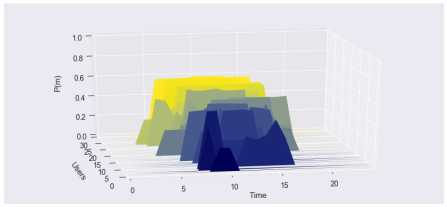
(j) Posterior means Bus/Outdoors and Recreation



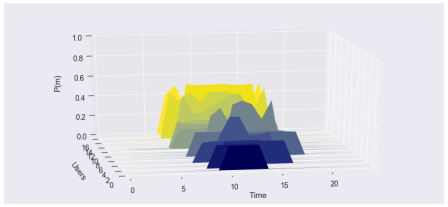
(k) Posterior means Rail/Outdoors and Recreation



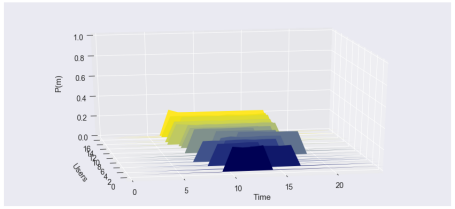
(l) Posterior means Tram/Outdoors and Recreation



(m) Posterior means Bus/Employment



(n) Posterior means Rail/Employment



(o) Posterior means Tram/Employment

Figure 7.7: Posterior means for the over sixty years old population group

Figures 7.8 and 7.9 below show aggregated boxplots of the transportation mode posterior distributions for all individuals in the three population groups, categorised by weekdays and weekends. It can be seen that the overall use of public transport for activity *Employment* is generally lower in the weekends for the unconstrained and > 60 group, particularly for using the bus. For the low income group, using rail for reaching activity *Eating and Drinking* is lower in the weekends compared to weekdays. On the other hand, weekdays dominate the use of public transport to reach *Education and Health* for the > 60 population group.

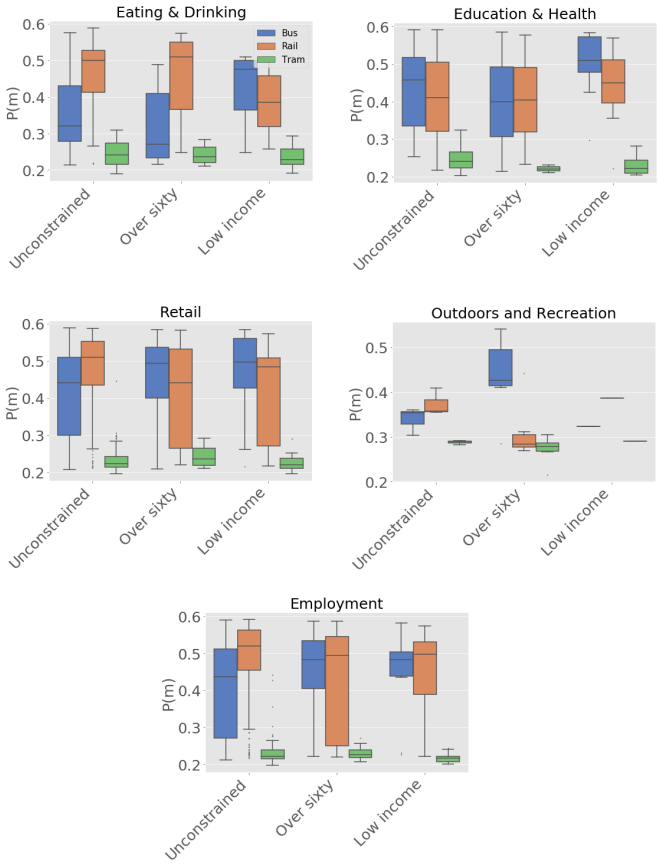


Figure 7.8: Aggregated transportation mode boxplots (weekdays)

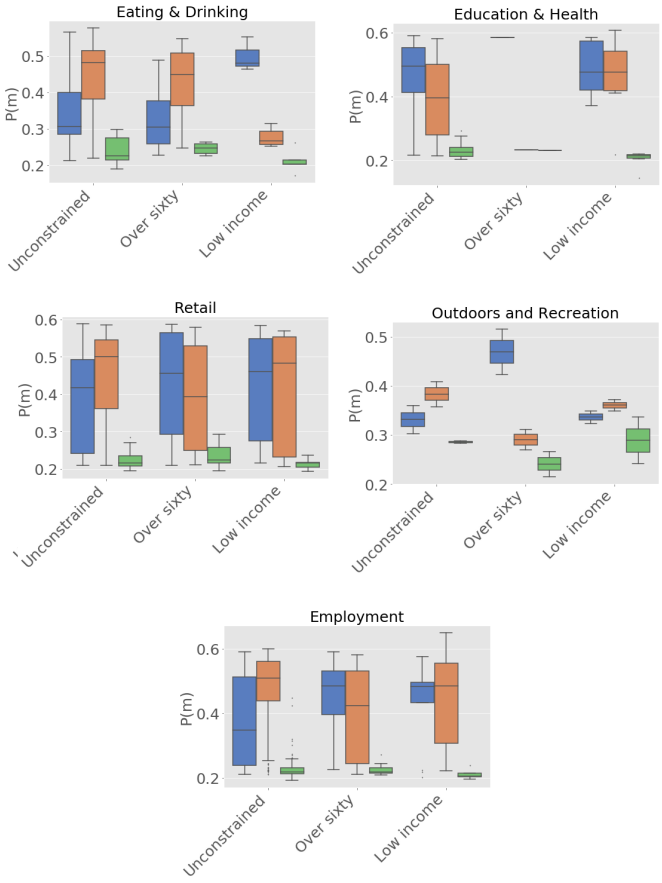


Figure 7.9: Aggregated transportation mode boxplots (weekends)

The findings for this posterior quantity are summarised below:

- Accessing *Eating and Drinking* activities by rail are peaking in late morning and afternoon/evening for the unconstrained group. Bus use is uniform.
- Rail is the dominant mode for *Retail* in the unconstrained group, peaking afternoon/evening.
- A grouping in the probabilities is observed between individuals of the unconstrained group for *Education and Health* for all transport modes.
- Accessing *Employment* using rail has slightly higher probability compared to the rest of the modes. For all modes, the commuting pattern to work is evident.
- Low probabilities for reaching *Eating and Drinking*, bus use dominates for low income group.
- Slightly higher probabilities for bus use for *Education and Health* activity for low income group.
- Double peak of *Employment* activities for low income group, with wider distributions for bus, rail.
- Use of bus and rail compliment each other for *Eating and Drinking* for the > 60 group.
- *Education and Health* activity shifted towards the afternoon hours with similar probabilities of bus and rail for the > 60 group.

7.2.3 Activity and mobility dynamics

Using individual trajectories from the passive mobility data, it is possible to evaluate the dynamics of activity types reached and transportation modes used for each individual. In the context of the CAA, these are represented by the transition matrices T_m and T_z are presented. Intuitively, these matrices capture the transition dynamics for the accessibility and mobility modules of the CAA model (figure 6.35), taking into consideration the effects of external factors as individuals transition from one transportation mode/activity to another during the trajectory. It is important to note that, contrary to T_m where the transportation mode states are inferred using the observed Oyster card modes, T_z captures the transition dynamics of inferred activity types.

7.2.3.1 Activity type transitions

Figure 7.10 below presents the posterior distributions for each element of the activity type transition matrix T_z for all individuals in the unconstrained population sample.

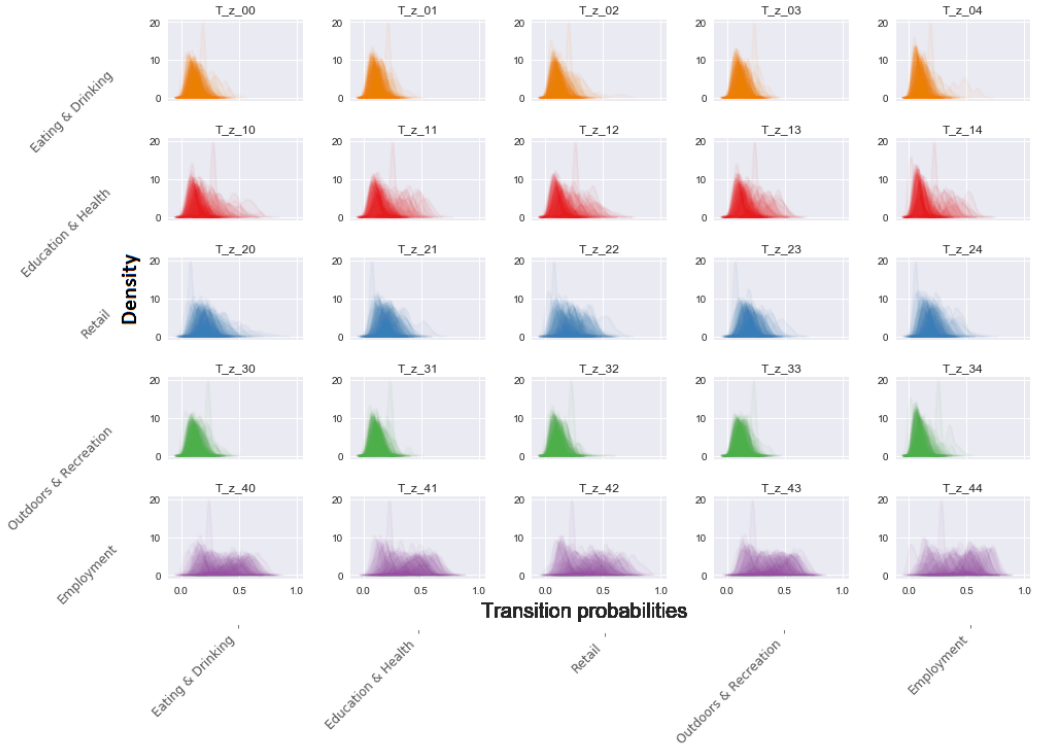


Figure 7.10: Posterior densities for T_z for the unconstrained population group

It can be seen that the transition patterns from activity type *Employment* to all other activities vary significantly between individuals, ranging from $0.2 < P(T_z) < 0.8$, with an overall mean probability of ≈ 0.4 for the transition from all other types to *Employment*, making this the dominant sequence in this group. Looking at the transition between *Education and Health* and *Education and Health*, individuals seem to be divided into two clusters, one with relatively low probability $P(T_{11}) < 0.2$ and one with probabilities $P(T_{11}) > 0.2$, a behaviour which could be attributed to the students/pupils in the sample. Relatively high probabilities for many individuals are also observed between transitions *Retail/Retail*, *Eating and Drinking/Retail*.

Results for the low income population group are shown in figure 7.11. The

transition patterns are very similar with the unconstrained population group. However, in this case the probabilities of transitioning from *Employment* to all other activity types is lower on sample population level $P(T_z) < 0.2$.

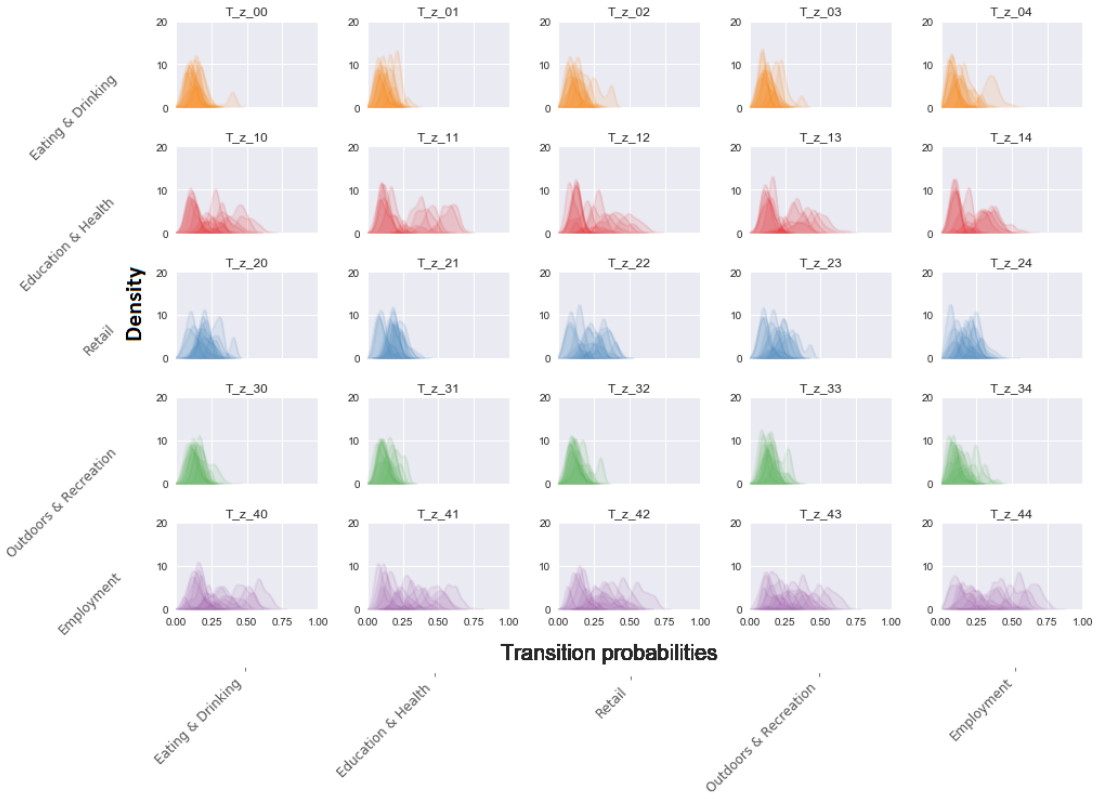


Figure 7.11: Posterior densities for T_z for the low income population group

Finally, figure 7.12 below presents the results for the > 60 years old population group. Again, the results here are very similar to the rest of the target groups, with the dominant transition sequences being between *Employment* and the rest of activity types.

The findings for this posterior quantity include:

- *Employment* related transition probabilities dominant with increased variability between individuals for the unconstrained group.
- *Education and Health* related transition probabilities appear clustered between individuals.

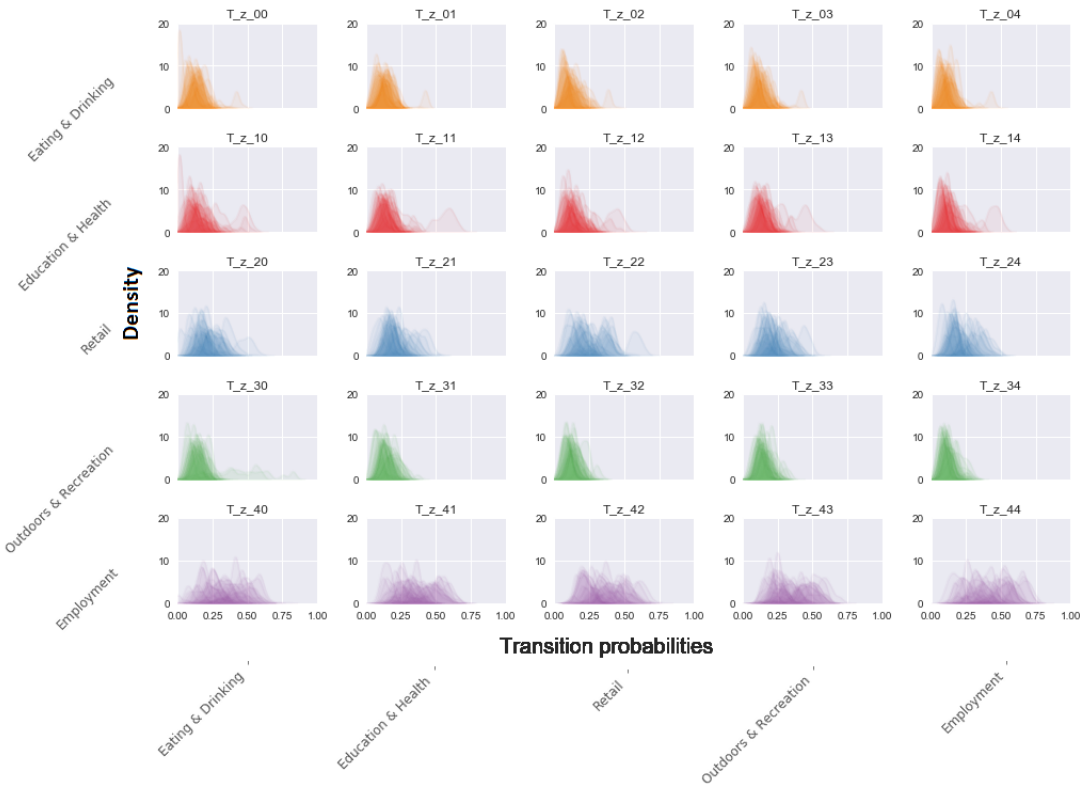


Figure 7.12: Posterior densities for T_z for the over sixty population group

- Low income group appear to have slightly lower *Employment* related transitions.
- >60 group transition probabilities are similar to the unconstrained group.

7.2.3.2 Transportation mode transitions

The posterior densities of T_m nodes are presented in figures 7.13, 7.14, 7.15 below for each element of the transition matrix.

Looking at the unconstrained population group, there is a clear tendency to persistently transition from rail services to rail services ($T_{m_{11}}$) with a population level probability of $P(T_{m_{11}}) \approx 0.65$, a behaviour which could largely be attributed to commuting to employment activities. The transition probabilities from bus to rail are also relatively high in the sample ($P(T_{m_{01}}) \approx 0.45$), on par with transition from bus to bus ($P(T_{m_{00}}) \approx 0.475$). The transition from rail to bus on the other hand ($P(T_{m_{10}}) \approx 0.305$) is comparatively low, indicating that, on average,

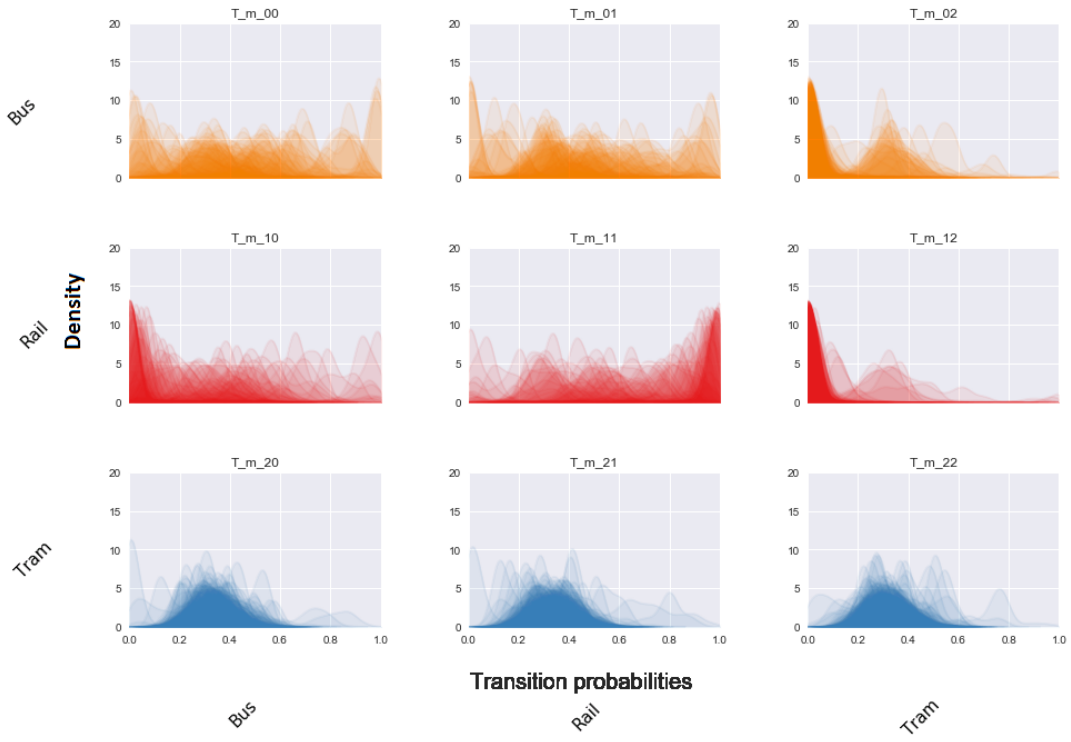


Figure 7.13: Posterior densities for T_m for the unconstrained population group

individuals of this sample seem not to prefer finishing their journey on the bus if rail was the prior choice. As expected, due to the lack of tram transactions in the sample but also due to the limited coverage of tram services, on average the transition probabilities between the rest of the transportation modes and tram is relatively small. This is confirmed by the uniform allocation of probabilities between tram and the rest of the modes. This pattern is similar to the over sixty and low income population groups.

Transition probability patterns for the > 60 population group (Figure 7.14) are different, providing evidence that, on average, there is an increased likelihood of using the bus persistently throughout a trajectory ($P(Tm_{00}) \approx 0.73$). The opposite is true for transitioning from bus to rail ($P(Tm_{01}) \approx 0.24$). Transitioning from rail to all other modes appears to be less clustered ($P(Tm_{11}) \approx 0.40, P(Tm_{10}) \approx 0.45$), indicating perhaps the less frequent use of rail services in this target group.

Finally, the low income population group (Figure 7.15) provides evidence of a broader spread of transition probabilities amongst individuals for the bus services, with a tendency to prefer using the bus throughout the trajectory ($P(Tm_{00}) \approx 0.55$) compared to transitioning from bus to rail ($P(Tm_{01}) \approx 0.41$). The overall

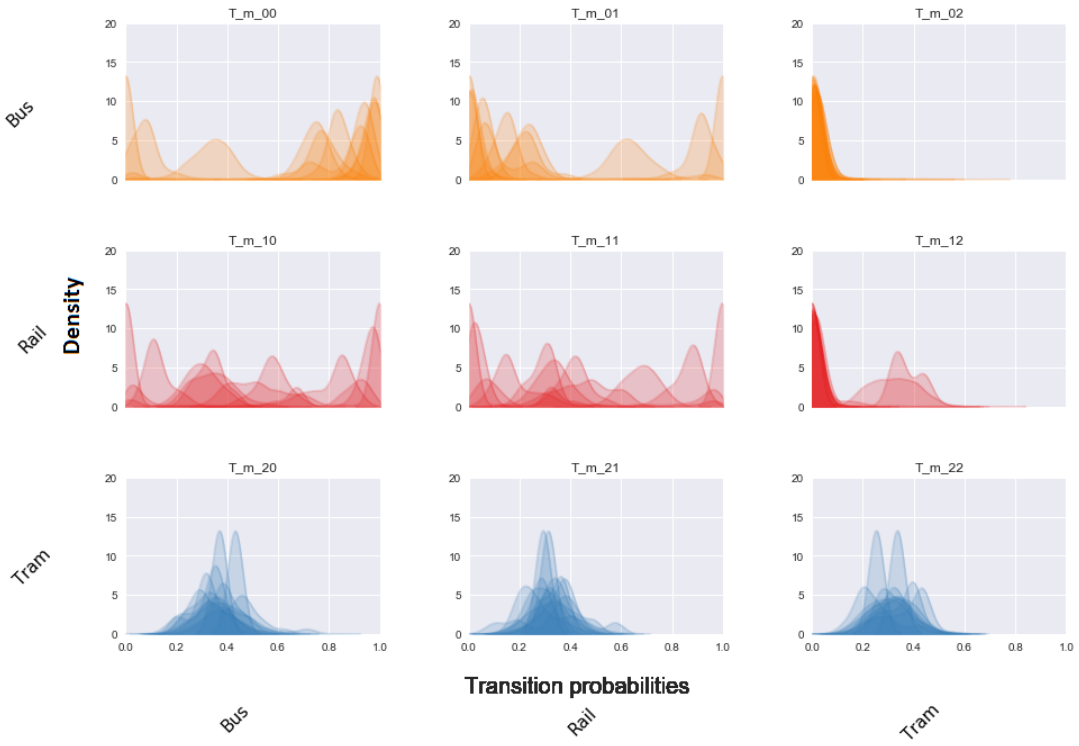


Figure 7.14: Posterior densities for T_m for the over sixty population group

pattern of rail use is similar to the > 60 population group, showing a uniform distribution of transitions between rail/bus and rail/rail ($P(Tm_{10}) \approx 0.46, P(Tm_{11}) \approx 0.43$).

The findings for this node can be summarised as:

- *rail/rail*, *rail/bus* transition probabilities higher, while *bus/rail* comparatively low for unconstrained group.
- *bus/bus* related transitions higher for > 60 and low income groups.
- *rail/bus* and *rail/rail* appear uniform for > 60 and low income groups.

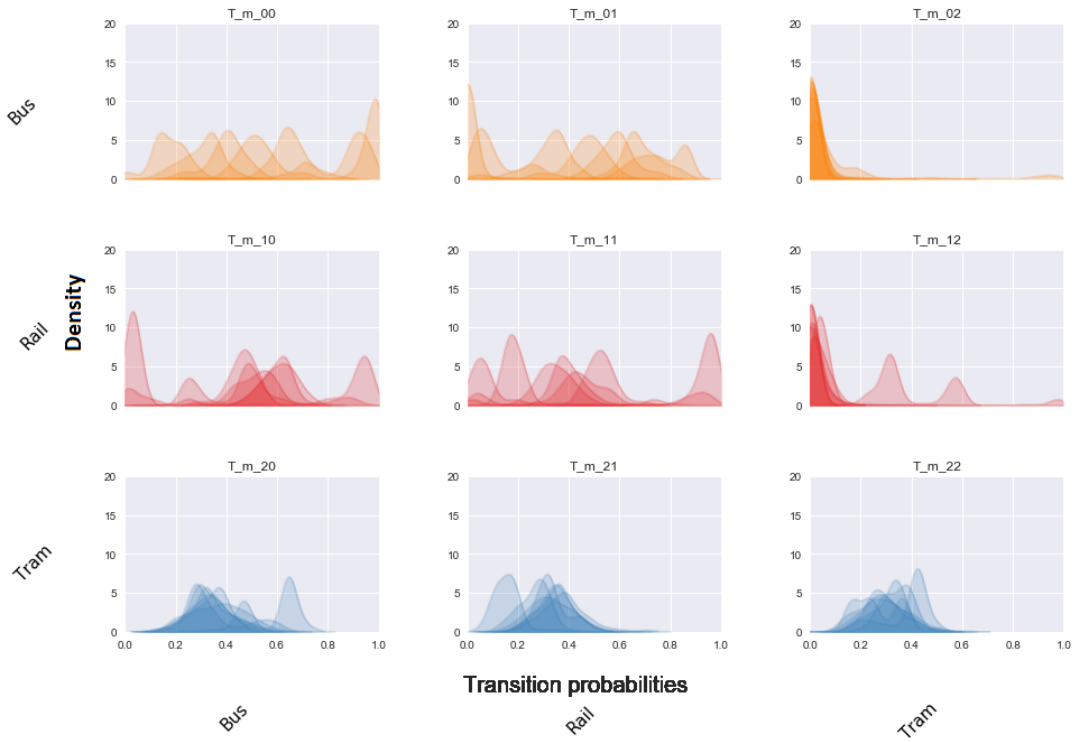


Figure 7.15: Posterior densities for T_m for the low income population group

7.2.4 Degree of contribution of external factors

External factors β that are assumed to influence an individuals' transportation mode/activity type dynamics are included in the modelling framework through the computation of transition matrices T_z, T_m. Each row of the transition matrices were modelled using a categorical distribution with probability derived by a softmax function (see section 6.4.4) on the place based covariates described in section 6.4.1.2. The categorical distributions were defined as sequences of state transitions from one state to the rest (section 6.4.2.3). Examining the posterior regression coefficients for each state allows one to assess the contributions of each factor on the transition probabilities from one mode/activity type to all others.

7.2.4.1 Activity types

The four external factors for transitioning from activity type κ to κ included in the accessibility module of the model in figure 6.35 were:

- $\beta_{\kappa,0}$: IMD
- $\beta_{\kappa,1}$: Proportion of green spaces
- $\beta_{\kappa,2}$: Traffic density
- $\beta_{\kappa,3}$: Crime rate

Figure 7.16 below shows boxplots of β posterior densities for each individual in the unconstrained population group. The colorbar maps to the mean value of the covariate effect.

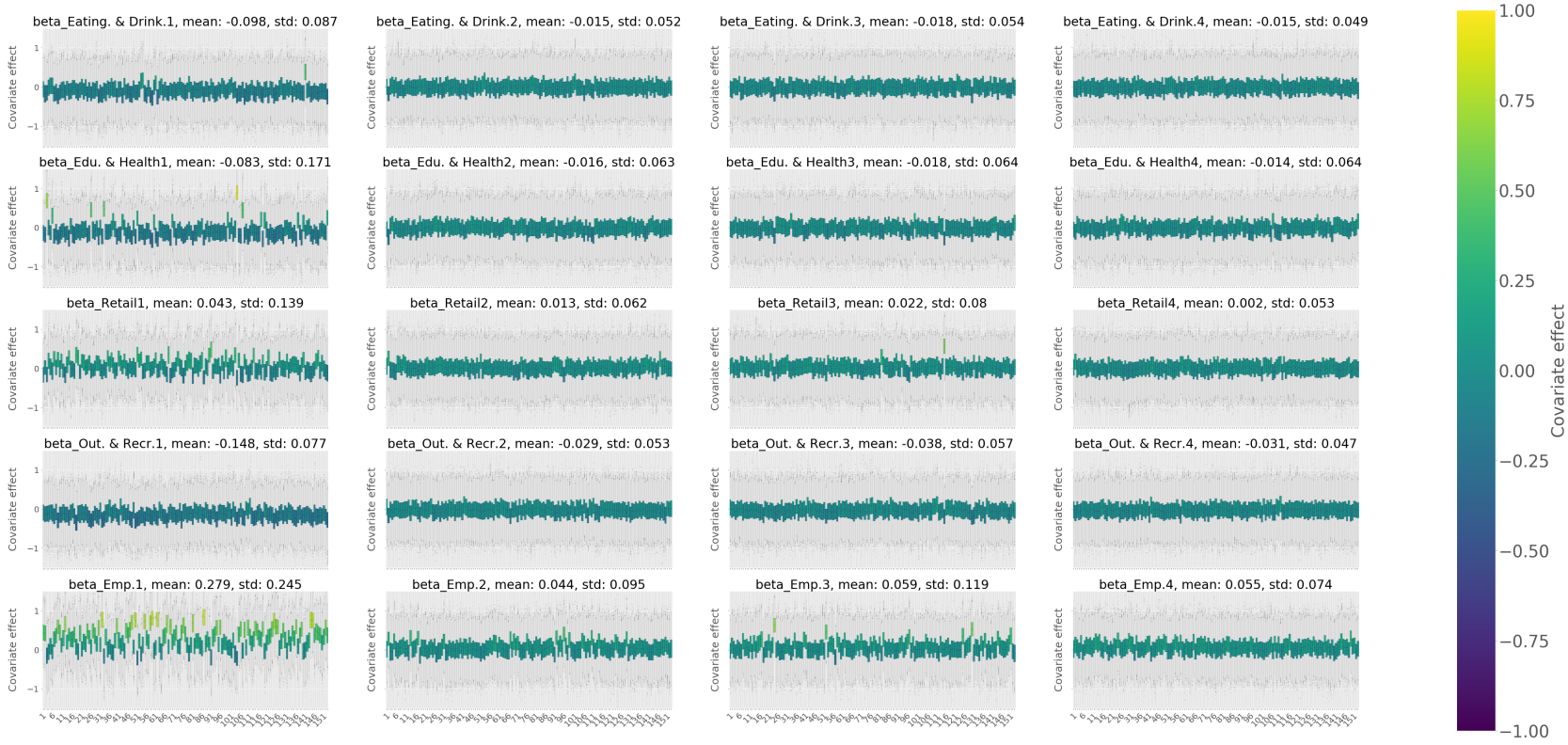


Figure 7.16: Posterior densities for *beta* for the unconstrained population group

Interpreting the results in the context of the model of equation 6.14, *IMD* seems to be the only covariate that has a weak but statistically significant effect for *Retail* and *Employment* related transitions for some individuals, given that the zero value is not contained within the Q1-Q3 interquartile range defined by the posterior samples. A positive sign in these coefficients signifies an increase in the chances of performing *Retail* and *Employment* related activity transitions as *IMD* increases. It is interesting to observe that for many individuals there is an inverse relationship in the *IMD* coefficient sign between *Retail* and *Employment*. This means that, for any transition sequence, a unit increase of *IMD* will increase the probability of one category and decrease the other depending on the coefficient sign. The rate of change between the probabilities varies for each individual and can be recovered using the softmax function on the inferred coefficients. As an example, increasing the *IMD* covariate by 0.1 for individual #25 (with coefficient vector $\beta_{IMD} = [-0.051, -0.059, -0.031, -0.085, 0.177]$) results in an increase in the probabilities of *Employment* related transitions by 0.01 and a decrease of 0.002 in *Retail* related transitions (assuming all other variables are constant). On the other hand, for individual #2 a same increase in *IMD* results in an increase for *Retail* by 0.005 and a decrease in *Employment* by 0.002 (with coefficient vector $\beta_{IMD} = [-0.022, 0.002, 0.18, -0.117, -0.060]$). An increase of activity type probabilities as a result of an increase in *IMD* may seem counter-intuitive, however, it should be taken into consideration that the trajectories for those individuals are characterised by repeated visits to areas of high *IMD* and, combined with the results of the rest of the CAA model's nodes, could be an indication of people experiencing higher risk of being spatially restricted to areas of high deprivation. It should be noted that this change is not constant and increases (or decreases) exponentially according to the softmax function.

For the remaining activity types (and the majority of individuals) the coefficients were found statistically non-significant given that the 0 value is contained within the interquantile range. This is not surprising since the majority of Oyster card transactions were related to *Retail* and *Employment* activity type transitions.

The β posterior densities for the over sixty years old population group are similar to the unconstrained population group with the most variation observed for the coefficients of *Retail* and *Employment* related transitions. Again here, the coefficient sign varies among individuals, with some having a decreasing effect in the probabilities of the transition sequence, and other individuals having an increasing effect (Figure 7.17). Similarly to the unconstrained group, the magnitude of the covariate effect for *Employment* related transitions is greater compared to *Retail*. As an example, an increase of the *IMD* covariate for individual #3 hav-

ing coefficient vector $\beta_{IMD} = [-0.108, -0.193, -0.004, -0.148, 0.425]$ results in an increase of *Employment* related transitions by 0.015.

Finally, Figure 7.18 presents the results for the low income population group. Similarly to the rest of the target groups, *IMD* seems to be the only covariate with a statistically significant effect for some individuals. However, contrary to the rest of the groups, the coefficient is significant for the *Education and Health* and *Employment* related activities. A possible reason for this could be the relatively large proportion of students in this group. To illustrate the effect of the coefficients to a unit change of *IMD* by 0.1, *Education and Health* transition probabilities increase by 0.25 for individual #8 with $\beta_{IMD} = [-0.247, 0.744, -0.18, -0.2, -0.121]$.

Figure 7.17: Posterior densities for *beta* for the over sixty years old population group

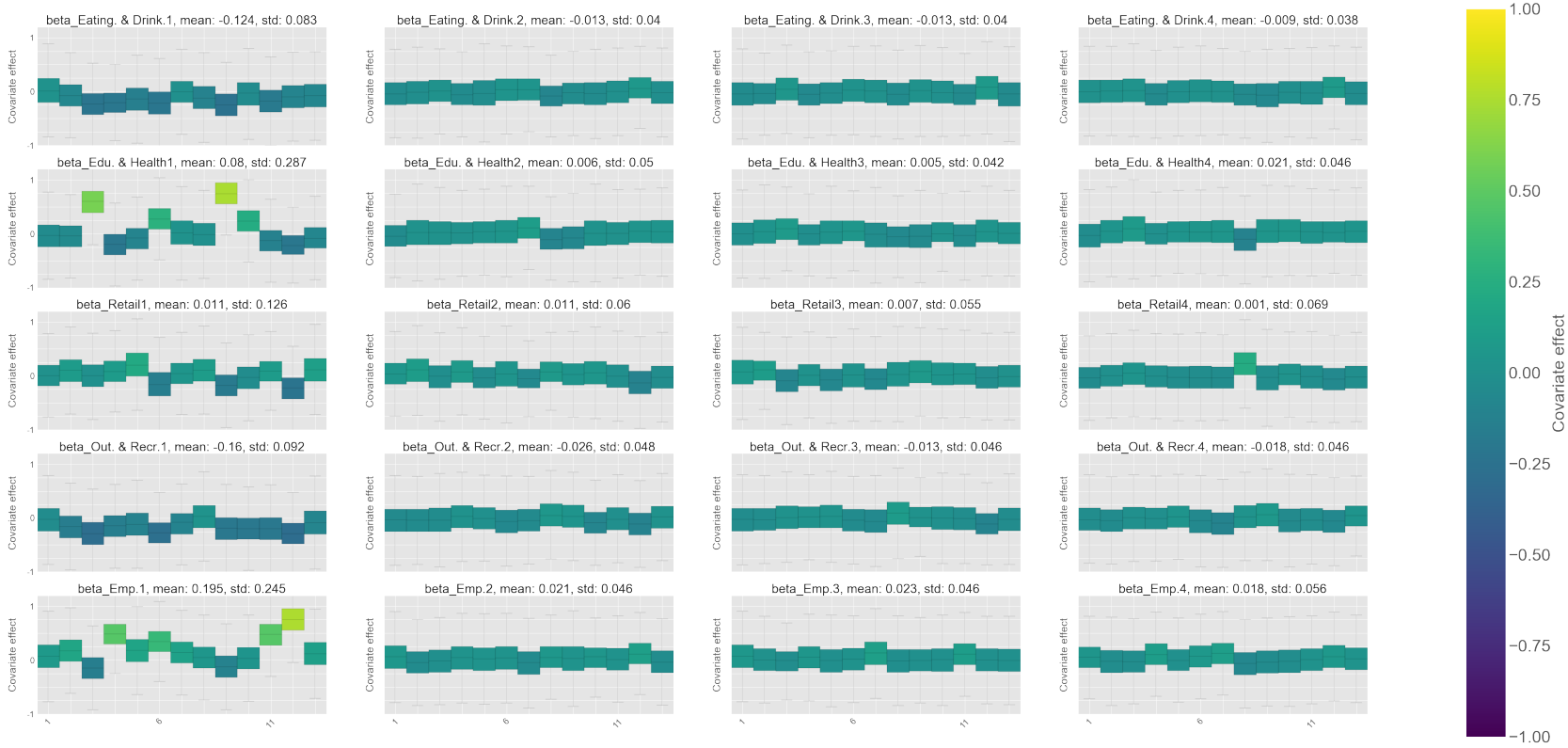


Figure 7.18: Posterior densities for *beta* for the low income population group

The above described findings can be summarised to:

- Covariate effects and signs vary across individuals.
- *IMD* covariate effect weak but statistically significant for *Retail* and *Employment* for may individuals in the unconstrained and > 60 population groups.
- The *IMD* effect is larger for *Employment*.
- *IMD* covariate effect is weak but statistically significant for *Education and Health* and *Employment* for some individuals in the low income group.

7.2.4.2 Transportation modes

The external factors for all transitions using transportation mode m to mode m included in the mobility module of model 6.35 were:

- $\beta_{m,1}$: Trip duration
- $\beta_{m,2}$: $PTAL$ index
- $\beta_{m,3}$: *xhr* Cyclic variation covariate ($sin(2\pi/24)$) for time of day
- $\beta_{m,4}$: *yhr* Cyclic variation covariate ($cos(2\pi/24)$) for time of day

Looking at the results for the unconstrained sample in figure 7.19 below, both *Trip duration* and $PTAL$ covariates seem to have a statistically significant effect (positive or negative) for all three transportation mode transitions for many of the individuals in the sample. For the majority of individuals of this population group, a unit increase of the $PTAL$ covariate results in an increase in the probability of *Rail* related transitions. Intuitively, this makes sense since higher $PTAL$ values are associated with a better coverage transportation network. For example, for individual #3 in the sample, an increase of $PTAL$ by 1 results in an increase of *Rail* probabilities by 0.03 and a decrease in *Bus* related transition probabilities by 0.027.

The results of the over sixty population group are shown in figure 7.20. Similarly to the unconstrained population group, both $PTAL$ and *Trip duration* covariates have a statistically significant effect for *Bus* and *Rail* related transitions. For the majority of individuals in this group, a unit increase of *Trip duration* results in an increase of probability of bus related transitions accompanied with a decrease in rail transition probabilities, while the inverse holds true for a unit

increase of *PTAL*. As an example, an increase of *Trip duration* by 1 minute results in an increase of *Bus* related transitions by 0.0002 for individual #8 with coefficient vector $\beta_{dur} = [0.21, 0.20, 0.23]$. On the other hand, a unit increase of *PTAL* covariate for this particular individual $\beta_{PTAL} = [0.12, 0.27, -0.39]$ results in a decrease of *Bus* transition probabilities by 0.024 and an increase of *Rail* by 0.06.

Finally, the posterior regression coefficients for the low income population group are presented in figure 7.21. Again, both *PTAL* and *Trip duration* covariates were found to have a statistically significant effect for some individuals in the sample. However, contrary to the unconstrained group, the magnitude of coefficients for *Bus* related transitions appears to be larger, driven by the bus dominated transport mode sequences for this target group. Depending on the nature of these transitions, for some individuals a unit increase of *PTAL* results in an increase of *Rail* over *Bus* related transitions while a unit increase of *Trip duration* results in an increase of *Bus* over *Rail* (e.g. individual #7).

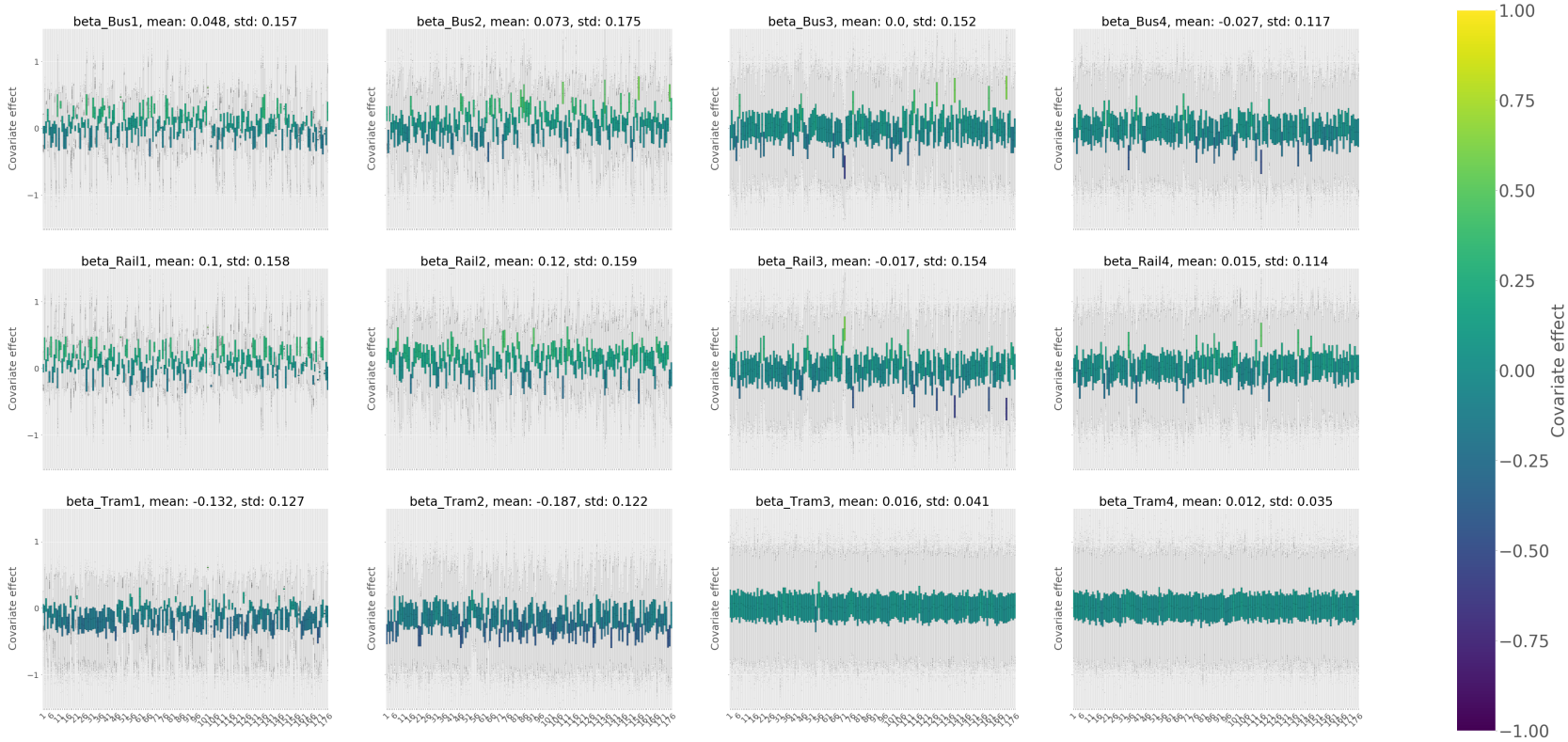


Figure 7.19: Posterior densities for *beta* for the unconstrained population group

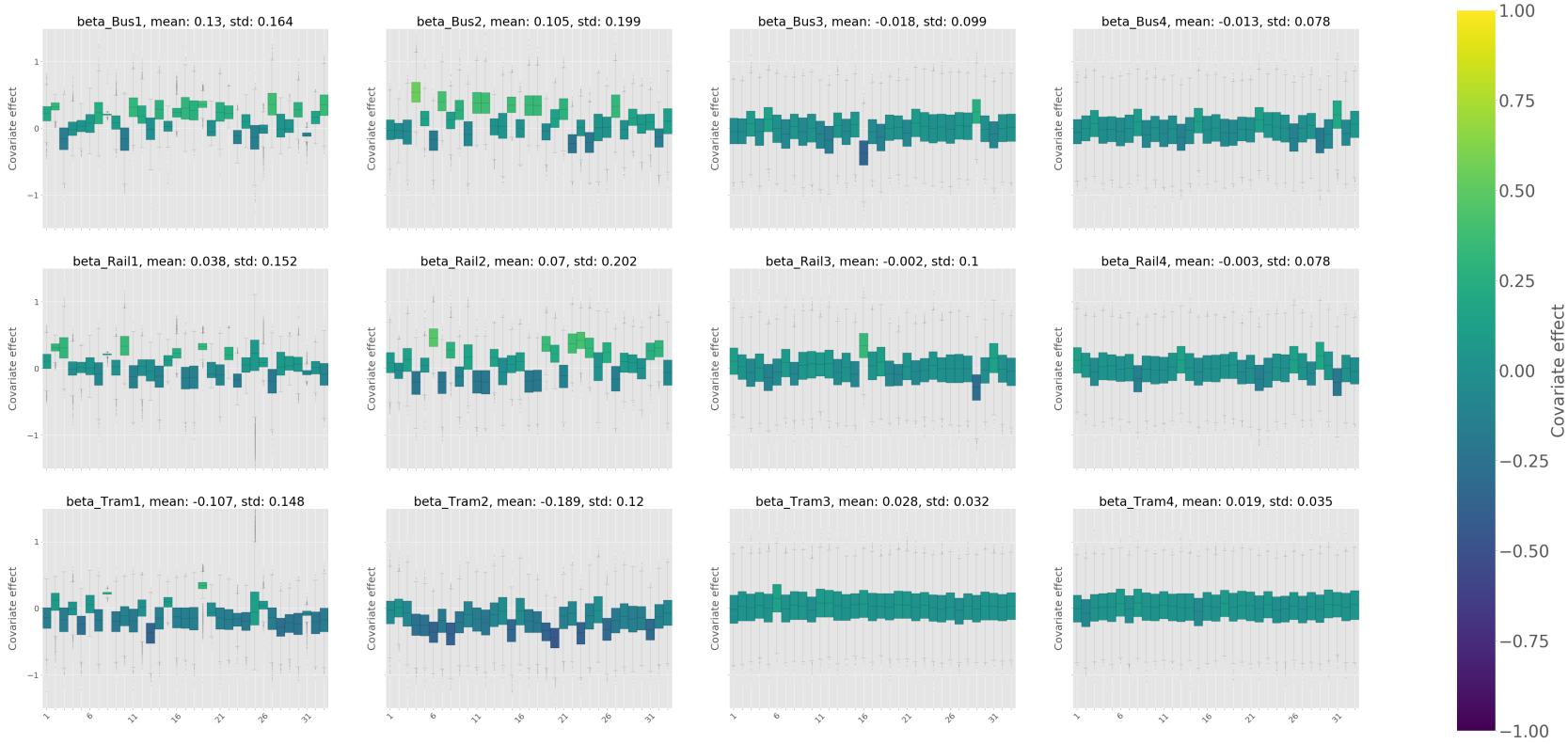


Figure 7.20: Posterior densities for *beta* for the over sixty population group

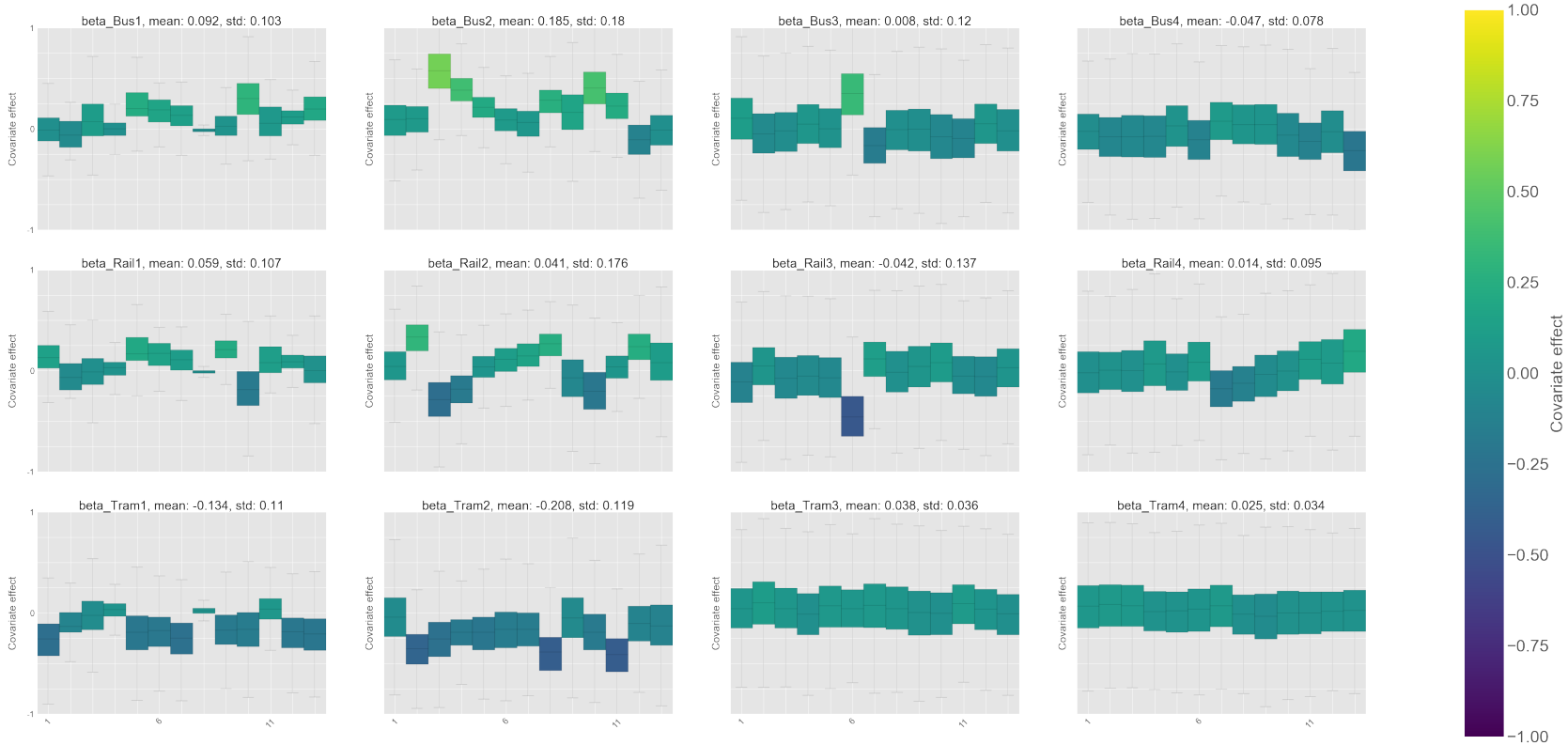


Figure 7.21: Posterior densities for β_m for the low income population group

Summarising, the findings for this node include:

- *PTAL* and *Trip duration* are statistically significant for many individuals in all groups.
- The magnitude of coefficients for *bus* higher for the low income group.

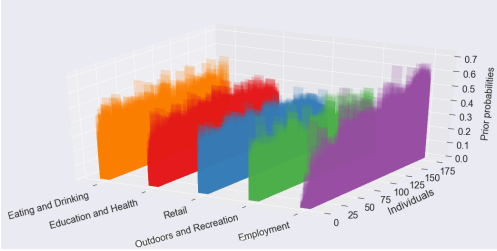
7.2.5 Degree of influence of sociodemographic characteristics

Sociodemographic characteristics were included in the CAA model shown in figure 6.35 as prior distributions derived from fitting a set of multinomial regressions on external data, namely the LTDS and RODS dataset (section 6.4.1.1). While the primary purpose of including such information in the model is to assist the task of activity type/transportation mode inference using the unlabelled Oyster card dataset, examining the final posterior distributions of the priors after observing the Oyster card transactions can reveal the degree of alignment between any prior assumptions on the levels of accessibility/mobility and the data. For the CAA model, the prior assumptions were incorporated through modelling the concentration parameters of the Dirichlet prior using the predicted probabilities as the mean parameter of a set of truncated normal distributions.

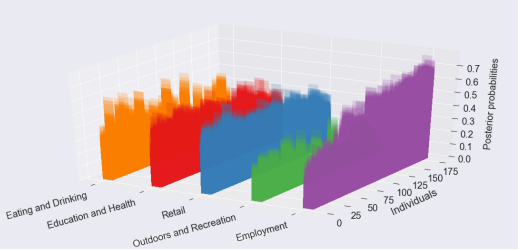
7.2.5.1 Activity types

Figure 7.22 below shows the prior and posterior distributions for the concentration parameters of activity types for all three population groups. For visualising the prior, random samples were generated for each activity space/individual throughout the day derived from RODS predicted probabilities using sex, age and arrival time as predictors. To visualise the resulting probabilities of the concentration parameters, a total number of 100 samples were drawn from a Dirichlet distribution using the posterior α_z. Note that in all cases, the individuals were sorted by increasing *Employment* probabilities.

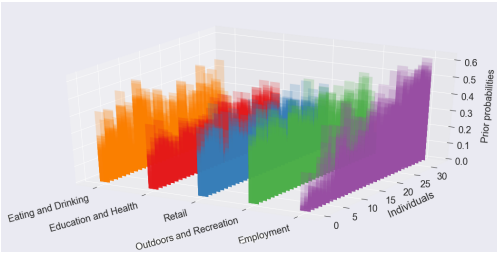
Looking at the unconstrained population group of figure 7.22 one could see that, overall, the shape of Dirichlet posterior follows the prior for the majority of individuals. This is less pronounced for the over sixty and low income population group, with the posterior estimates moving away from the prior for the majority of individuals. This is not surprising, as the predicted probabilities derived by RODS regression do not take into account variables such as income or employment status. In all population groups, the posterior Dirichlet for activity type *Outdoors and Recreation* has moved away from the prior, reflecting the lack of this particular activity type in the inferred Oyster card data.



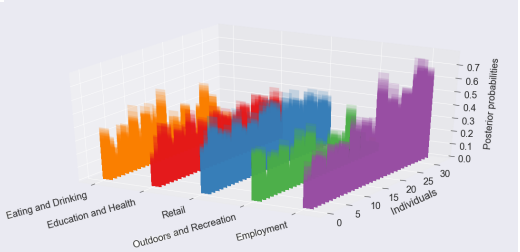
(a) Prior Dirichlet samples for the unconstrained Oyster group.



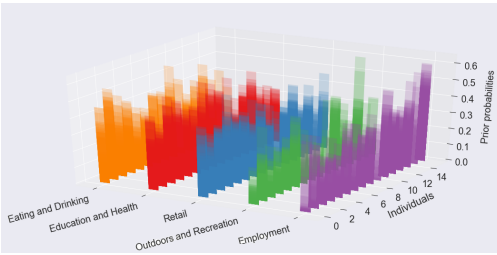
(b) Posterior Dirichlet samples for the unconstrained Oyster group.



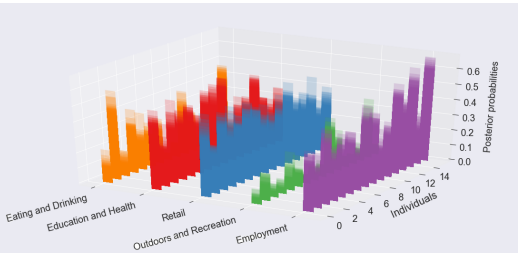
(c) Prior Dirichlet samples for the over sixty population group.



(d) Posterior Dirichlet samples for the over sixty population group.



(e) Prior Dirichlet samples for the low income population group.



(f) Posterior Dirichlet samples for the low income population group.

Figure 7.22: Prior and posterior Dirichlet samples for the different population groups.

Summarising:

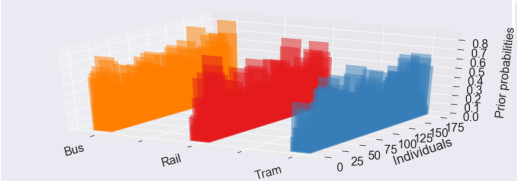
- Prior/posterior distributions for the effect of sociodemographic characteristics are in line for unconstrained group.
- Posterior distributions different from prior for the > 60 and low income group.

7.2.5.2 Transportation modes

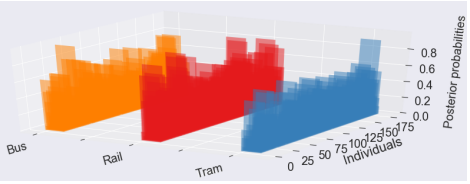
The prior/posterior probability distributions for the mobility part of the CAA model (figure 6.35) is shown in figure 7.23 below. Contrary to activity types, the

predicted probabilities in this case were generated through fitting three independent binary logistic regressions on the LTDS dataset, this time using a wider set of sociodemographic variables. Due to the absence of arrival time as a covariate, the resulting transportation mode prior distributions persisted throughout an individual's trajectory.

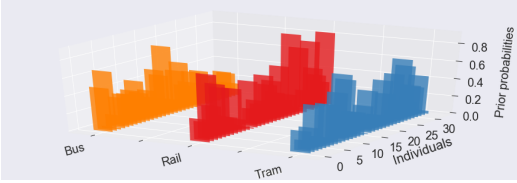
Similar to the case of activity types, the prior/posterior relationship of Dirichlet probability samples for the unconstrained population group appear to be consistent, with the difference of a probability reduction for *Tram* services fluctuating around the random 0.33 probability threshold. On the other hand, the posterior Dirichlet samples for the individuals on the remaining population groups deviate from the prior assumptions for nearly all Oyster card users. In particular, the posterior proportion of transportation modes for the > 60 population group was found to have higher probabilities for *Bus* services. The prior assumptions for the low income population group have also been updated after observing the Oyster card trajectories. For this population group, the use of *Rail* services was updated with higher probabilities together with a reduction in the corresponding *Bus* use for a small portion of the sample. Furthermore, the prior peaks of *Tram* use have also been smoothed, appearing to complement the *Bus* probabilities.



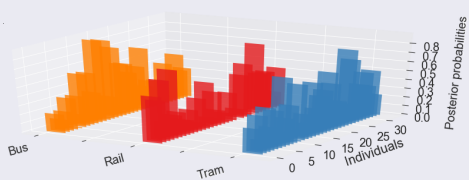
(a) Prior Dirichlet samples for the unconstrained Oyster group.



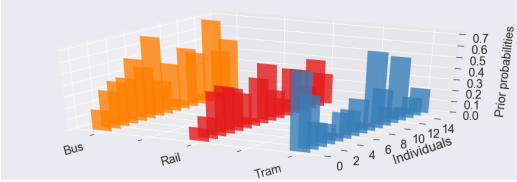
(b) Posterior Dirichlet samples for the unconstrained Oyster group.



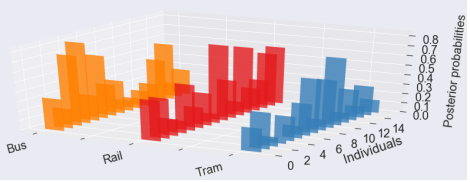
(c) Prior Dirichlet samples for the over sixty population group.



(d) Posterior Dirichlet samples for the over sixty population group.



(e) Prior Dirichlet samples for the low income population group.



(f) Posterior Dirichlet samples for the low income population group.

Figure 7.23: Mobility prior and posterior Dirichlet samples for the different population groups.

Summarising:

- Results show a prior/posterior consistency for the effect of sociodemographic attributes for the unconstrained group.
- For the > 60 group, prior assumptions related to the effect of sociodemographic attributes for *bus* use have been increased.
- For the low income group, prior assumptions related to the effect of sociodemographic attributes for *rail* use have been slightly increased.

7.2.6 Results summary

In this section, the posterior quantities defined in the model of section 6.4 were presented for each of the models' nodes. The results revealed a number of distinct accessibility patterns which could be associated with different qualitative characteristics of the three population groups.

In particular, the probabilities of the low income group for activity types *Eating and Drinking* and *Retail* were found to be considerably smaller compared to the > 60 and unconstrained target group, providing evidence of reduced access to these activity types. Moreover, the temporal patterns of *Employment* activity type was found to vary between the three target groups, indicating different levels of flexibility when reaching for this activity type. The probabilities of transitioning from *Employment* to all other activity types was also found to be larger in the unconstrained and > 60 group compared to the low income group. In terms of the transportation modes used to reach the different activity types, the three target groups are characterised by different proportions, with the unconstrained group appearing to be *Rail* dominated. On the other hand, the low income group appears to make use of the *Bus* more, while for the > 60 group the distribution is more balanced. These results are also reflected in the posterior distributions of the transportation mode transition matrix, with probabilities of transitioning from *Rail* to all other modes being significantly larger compared to the rest of target groups.

In terms of external factors used to model the individual activity type transitions, only *IMD* was found to have a weak but statistically significant effect for all target population groups, with *Retail* and *Employment* related transitions being affected the most from a change of this covariate. For transportation mode related covariates, both *Trip duration* and *PTAL* were found to have a significant influence in *Bus* and *Rail* related transitions for the majority of individuals in the sample.

Finally, the relationship between the prior before and after observing the data for activity type inference, revealed that the individual characteristics such as income and age play an important role in shaping the distributions of different activity types. This was more evident for the individuals that have greater discrepancy between the prior before and the posterior after data were observed. For the low income group, this discrepancy was even greater for all defined available modes.

These results will be analysed further in the next section, which will seek to examine the within and between population group variations in posterior densities, using cross entropy as a tool to provide evidence of social exclusion and transport disadvantage experienced by some individuals.

7.3 Evaluating individual based social exclusion using the Capabilities approach to accessibility model

In this section, the results of section 7.2 are evaluated with regards to assessing relative transport disadvantage and social exclusion between the unconstrained, > 60 and low income population group. For this task, a popular entropy based equality index is applied to the posterior quantities of each member of the capabilities set. Finally a discussion of the posterior results in relation to the distribution of the individual equality indices is offered, aiming to highlight aspects of social exclusion and transport disadvantage experienced by some individuals. The analysis and discussion of this section has been published in the article "Assessing transport related social exclusion using a Capabilities Approach to accessibility framework: A dynamic Bayesian network approach" of journal *Journal of Transport Geography* (Bantis & Haworth 2020).

7.3.1 Defining an accessibility assessment framework within the context of social exclusion

Regarding the definition of an accessibility comparison framework, some authors (Van Wee & Geurs 2011) argue that specification of a minimum threshold under which lack of accessibility results in increased social exclusion requires a degree of moral judgment, and as such is primarily a political choice dependent on the history and values of a particular society. Others (Pereira et al. 2017) argue that it is the responsibility of institutions and policy makers to assess the impact on equality caused by an intervention on the wider population, and how this can be addressed. In any case, the problem of setting a minimum acceptable accessibility threshold remains an open challenge (Hananel & Berechman 2016, Farrington & Farrington 2005). To that end, literature focused on assessing the social outcomes of transport interventions seems to favor theories that focus on relative differences between individuals/groups of individuals rather than absolute levels of accessibility (Lucas et al. 2016, El-Geneidy, Buliung, Diab, van Lierop, Langlois & Legrain 2016). The premise is that the least advantaged members of society should benefit the most from any interventions aimed at improving accessibility. Even in this case, however, moral judgment by decision makers on what constitutes an acceptable range of differences is inevitable. Nevertheless, such an approach is a clear step towards identifying the factors that contribute the most to different levels of accessibility, which can then guide policy makers to specific decisions that will maximise accessibility benefits for the least advantaged groups.

Within transportation literature, comparison of different levels of accessibility outcomes has been approached in different ways (Geurs et al. 2016). The first involves comparing accessibility levels of individuals belonging to the same group of needs (horizontal equity). These groups can be people from the same socioeconomic backgrounds, gender, ability etc. The second refers to the comparison of individuals across different backgrounds (vertical equity). The third compares accessibility levels in the spatial domain (spatial equity) and finally, the fourth compares population groups with different needs and abilities (social equity). In terms of tools used to quantify accessibility outcomes from an equity perspective, researchers have used different statistical measures, ranging from simple ones (such as the variance or the range) to more complex ones such as the Atkinson and Kolm inequality measures (Ramjerdi 2006). It is important to note that these indices benchmark the comparison against an idealised state, and from this perspective they measure equality. To derive arguments related to equity, an empirical interpretation of results is required.

A commonly used measure is the Gini index (Geurs & Ritsema van Eck 2001, Neutens, Schwanen, Witlox & De Maeyer 2010, Delbosc & Currie 2011*b*). This is a statistical dispersion measure that uses the curve produced by the cumulative distribution function of one variable (typically an accessibility measure, income, service supply etc.) and of the proportion of ordered population. A commonly reported disadvantage of the Gini index is the lack of decomposability. This means that the index is not easily decomposed for different population groups, so that the sum of the different components equals the total amount of the Gini index (World Bank Group 2005). In this regard, the Gini index can be considered as a measure of horizontal equity (Camporeale et al. 2017).

Another statistical measure used to quantify accessibility outcomes from an equity perspective is the Theil index (Delafontaine et al. 2011, Santos et al. 2008). Its formulation is based on information theory's definition of entropy, with the difference that it uses the base of natural logarithm and the accessibility measures used are normalised by the population mean. Theil's index satisfies most of the desirable properties of an equality measure (World Bank Group 2005), such as being scale invariant, population independent and decomposable. It also satisfies the Pigou-Dalton Transfer sensitivity, which states that transfer of benefits from those who are better-off to those who are not reduces inequality levels. Another reported benefit is computational efficiency since, contrary to the Gini index, Theil's index can be computed in linear time (Delafontaine et al. 2011). In addition, since the index is based on the theoretical properties of entropy, it is possible to be modified to be used within other information theoretic entropy based frameworks such as

relative entropy between individuals or population groups (Roberto 2015).

Entropy based measures are attractive for the task of assessing equality levels of individuals within and between population groups within the CA based accessibility framework. This is because entropy based indices can be directly applied to the output posterior probability distributions of the Bayesian network for each individual, providing information on the levels of diversity between the probability distributions of nodes. The results can then be used to provide evidence on reduced access to activity types/public transportation modes.

7.3.2 A Theil index based assessment framework

Within the developed framework of CAA, two components of an individual's ability to reach opportunities were identified and quantified, given personal characteristics and external factors: 1) potential accessibility to different activity types using the public transport, and 2) potential mobility of using public transportation modes. The first one is related to the concept of equality of reaching opportunities while the second is related to issues of transport disadvantage. The next step of the analysis is to further explore these components using the posterior quantities as a basis of comparison.

As already noted, entropy based measures have been used within the context of measuring equity, with the Theil index being an example. Within the wider accessibility literature, index has been proposed as theoretically capable of quantifying accessibility related equity issues (Van Wee & Geurs 2011) and has been applied as an equity evaluation tool for different case studies. Examples include equity evaluation under different configurations of public service opening hours (Delafontaine et al. 2011) and developing performance indicators of accessibility measures to quantify cohesion effects of transport infrastructure investments (López et al. 2008).

The Theil index quantifies the actual entropy relative to the maximum entropy of the data and practically is a measure of the difference between complete randomness and uncertainty and the observed state of the dataset (equation 7.1):

$$S_{Theil} = \sum_{i=0}^{N} \left(\frac{x_i}{N\bar{x}} ln \frac{N\bar{x}}{x_i} \right) \tag{7.1}$$

$$S_{max} = lnN$$

$$T = S_{max} - S_{Theil}$$

where x is a vector of non-negative elements, S_{Theil} is the observed entropy and S_{max} is the theoretical maximum entropy of the dataset.

It is interesting to observe that the above formulation is similar to Kullback-Leibler (KL) divergence (or relative entropy) if the vector x is a valid discrete probability distribution, as is the case of the posterior distributions of the CAA model figure 6.35, and S_{max} is the maximum entropy defined by the cardinality of the event set. Kullback-Leibler divergence is a common probability divergence measure used to compare probability distributions within the context of applications in information theory (Cohen & Kempermann 1998) and statistics (Pardo 2005).

By definition, $T \geq 0$, with 0 meaning that the distribution is identical to the uniform distribution (the observed entropy is equal to the maximum) and higher values signifying increased deviation from the uniform case and thus increased inequality. It is important to note that the index is invariant under state switching in the set. For example two individuals, one using the bus 90% and the remaining modes 5% of the time, and a second individual using the rail 90% and the remaining modes 5% of the time, will be assigned the same Theil value. This doesn't take into consideration which transportation mode is more favourable under a given circumstance. From this perspective, arguments related to equity are not possible by assessing the output of the index alone, and some qualitative discussion of results is needed. This is also true for the weighted version of the index, as the weighting scheme needs to be decided to reflect equity considerations. Moreover, the uniform level of equality specified by maximum entropy represents a theoretical case that links to egalitarian approaches under the idea of equality of opportunity (Pereira et al. 2017). However, as has already been acknowledged in section 2.5, it is legitimate to expect a certain level of inequality to exist, provided that they are caused by an individual's own choices and not unfavourable circumstances such as having low income.

For the purposes of identifying individuals that experience a relative disadvantage, the posterior distributions will be compared and contrasted using the Theil index against the state of complete equality characterised by maximum entropy. Since the Oyster card dataset doesn't provide any information related to an individual's preferences or desires, Theil values will be assessed under the assumption that any significant deviations of the individual Theil values from the group population mean could be attributed to particularities of the group (eg. low income, age), treating individual preferences as random fluctuation in the Theil values within the group.

7.3.2.1 Potential accessibility to activities

This element of the capabilities set corresponds to the (weighted) distribution of activity types that are assumed to be reachable within 5/10/15/20 minutes walking time from a transportation access point (see section 5.5.3.2). The equality assumption made here is that, throughout an individual's trajectory, all defined activity types should be equally reachable by an individual regardless of factors such as age, income etc., and thus the distribution of these activity types should approach the uniform distribution. A deviation from this hypothetical scenario is assumed to trigger issues of social exclusion as a certain level of options to participate in activities is not available given the trajectory.

Figure 7.24 below shows density plots of Theil indices for the posterior activity type distributions for each population group. The Theil values in this plot were calculated using the mean distribution over each individual's trajectory sequence, resulting in one index for each individual:

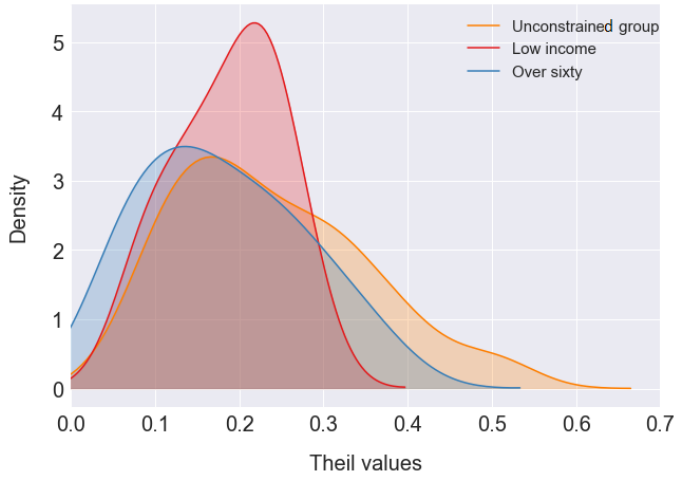


Figure 7.24: Density plots of Theil indices for the three population groups

Using a one way ANOVA test (table 7.2), the difference between the means between all three population groups were found to be statistically significant at the .05 significance level ($F - value = 4.424, p = 0.013$), indicating that they belong to different distributions (rejecting the null hypothesis).

The low income group has the largest mean compared to the rest of the groups, signalling overall increased inequality levels. The distribution of Theil indices for the over sixty and unconstrained population groups are similar. However, the tail of the unconstrained population group is considerably longer compared to both remaining groups. Under closer examination, these outliers ($> 75\%$ percentile, $T > 0.45$) are characterised by high *Employment* probabilities (> 0.6) with *Shopping*

group	count	mean	std.	75 perc.
Unconstrained	181	0.18	0.11	0.31
Over sixty	30	0.12	0.09	0.25
Low income	13	0.21	0.06	0.23

(a) Descriptive statistics

	sum. sq.	df	F	p-value
group	0.105	2.0	4.42	0.013
Residual	2.62	221.0	NA	NA

(b) One way ANOVA

Table 7.2: Desciptive statistics and one way ANOVA for the Theil indices

activity type being second highest. The demographic status of these individuals is composed of a mix of ethnic statuses, while the place of residence is outer London in most cases. All of the individuals in this sample are in permanent full time employment in central London, with activity patterns being limited to <4-5 unique locations.

The outliers ($> 75\%$ percentile $T > 0.23$) of the low income group on the other hand, are characterised by individuals who are part-time workers and students, again residing in outer London. Their ethnic background is a mix of Asian/Arab/Black or Black British - African and Black or Black British - Caribbean with age spanning from 21 to 38 years old. The household characteristics are lone parents or couples with children. Compared to the outliers of the unconstrained group, the number of unique locations visited is greater. However, the mean distance between these locations (9.5km) is much smaller compared to the unconstrained group (20.4km). This pattern could be explained by the relatively high rates of travelling by bus and provides evidence of a reduced space where activities can take place compared to the unconstrained group. In the absence of access to individual preference mechanisms, it is difficult to make assertions as to whether this pattern is due to genuine individual choices or whether it is related to higher risk of social exclusion. However, given that in London the price of a single bus journey is nearly half the price of rail and taking into consideration the

sociodemographic profile of these individuals, it is likely that the observed pattern is due to necessity.

Finally, looking at the demographic characteristics of the outliers of the > 60 population group ($> 75\%$ percentile, $T > 0.24$), the ethnic backgrounds are mainly White - English/Welsh/Scottish/Northern Irish/Other White residing in outer Greater London with place of employment in Greater London area, inside the M25 motorway. All of the individuals in this percentile were employed full time with annual income ranging between £25,000-100,000. Similarly to the unconstrained population group, these individuals have high *Employment* activity type probabilities (~ 0.5). However, the probabilities for the rest of the activity types appear to be more balanced. This population subgroup has the greatest number of unique visited locations compared to the unconstrained and low income groups. However, contrary to the low income group, the mean distance between these locations is slightly larger (10km), a fact which could be explained by the higher rate of travelling by rail for activity *Eating and Drinking*.

7.3.2.2 Potential mobility

This element of the capabilities set expresses an individual's potential mobility through a probability distribution over the set of public transport modes used throughout a trajectory. The equality assumption here is similar to potential accessibility: all transportation modes should be equally available regardless of any personal or place based characteristics. It is important to note that this assumption is useful only in the context of benchmarking the individual Theil values, as it is well known that the public transportation network is designed so that each mode complements the other. Moreover, as in the case of *Tram* services in London, some transportation modes are operating on a local scale only, so by default are not readily available to the general population. Nevertheless, by evaluating the individual Theil indices in a relative way, it is possible to identify cases where the use of a transport mode is not possible due to factors beyond the control of an individual, a fact which could relate to transport disadvantage.

Figure 7.25 below shows density plots of Theil indices for the posterior transportation mode distributions for each population group:

Similar to Section 7.3.2.1, a one way ANOVA test was performed which resulted in failure to reject the null hypothesis, concluding that the distributions belong to the same population (table 7.4). However, this result could be an artefact of the lower cardinality of the transportation mode set, particularly considering the very low use of *Tram* services, resulting in small differences in Theil values.

Exploring the distributions qualitatively, one notices a bimodality in all three

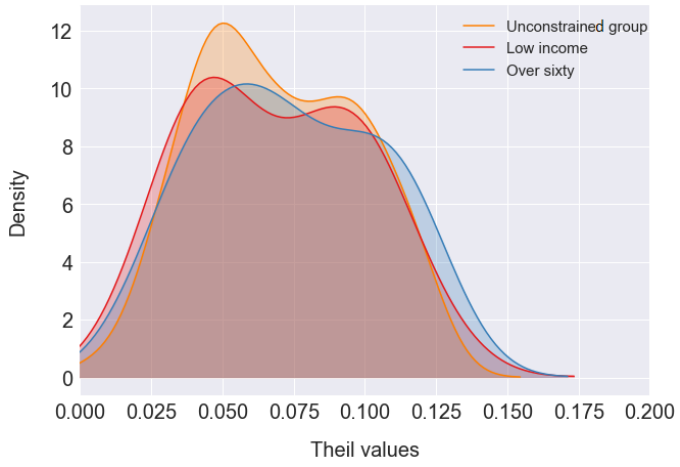


Figure 7.25: Density plots of mobility Theil indices for the three population groups

group	**count**	**mean**	**std.**	**75 perc.**
Unconstrained	181	0.07	0.03	0.1
Over sixty	30	0.073	0.029	0.11
Low income	13	0.06	0.02	0.09

(a) Descriptive statistics

	sum. sq.	**df**	**F**	**p-value**
group	0.00038	2.0	0.23	0.79
Residual	0.18	221.0	NA	NA

(b) One way ANOVA

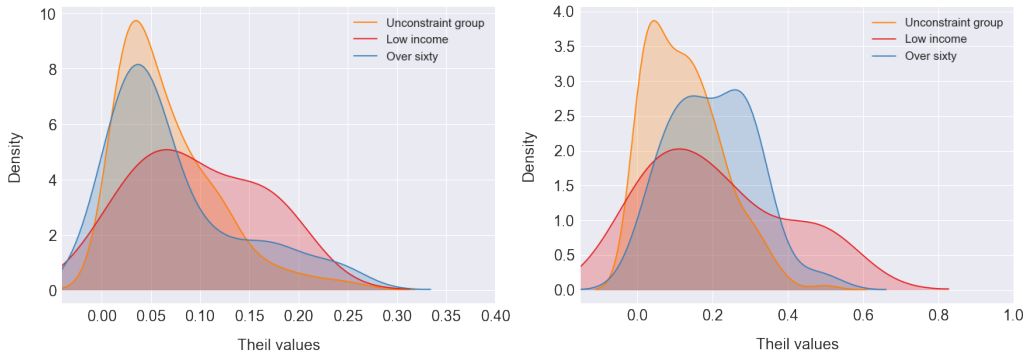
Table 7.4: Descriptive statistics and one way ANOVA for the Theil indices of transportation modes

population groups, meaning that, for those individuals, the use of one transportation mode dominates over all others. It is interesting to observe that, in contrast to the unconstrained population group, the second mode of the low income group is attributed to very high probabilities of *Bus* use. Examining the outliers ($> 75\%$ percentile, $T > 0.1$) of the low income distribution, one notices that the majority of individuals in this set are a subset of the low income outliers of section 7.3.2.1.

This fact provides further evidence of the potential for social exclusion for these individuals.

7.3.2.3 Public transport and activity dynamics

This element of the capabilities set aims to quantify the degree of interaction with activity types/transportation modes through the use of transition matrices informed by external factors. Regarding the interaction with activity types, the equality assumption here is that the an individual is less likely to be socially excluded if they maintain a uniform level of interaction with the available activities, as this translates to more frequent trips per activity type, a characteristic thought to map to increased levels of social involvement (Schönfelder & Axhausen 2003). A similar rationale holds for interaction with public transport modes as expressed through the mobility transition matrix, in that increased levels of transition between modes could translate to an expansion of the set of activities within reach. Figures 7.26a and 7.26b below show the Theil values between the three population groups for the activity type transition matrix.



(a) Density plots of Theil indices for the activity types transition matrix

(b) Density plots of Theil indices for the mobility transition matrix

Figure 7.26: Distributions of Theil indices for the transition matrices

The ANOVA test for the three population groups (table 7.5) failed to reject the null hypothesis (same distributions), however for the low income group there are some outliers that seem to have increased Theil values.

Similarly to 7.3.2.1, the employment status of outliers ($>$ 75% percentile, $T > 0.2$) of the low income group are a mixture of student, part-time and full time workers, residing in outer Greater London. Not surprisingly, these individuals are characterised by increased probabilities of transitions related to *Health and Education* (for the student and part-time employed individual) and increased transition probabilities related to *Employment* for the full time workers. The lat-

group	count	mean	std.	75 perc.
Unconstrained	181	0.06	0.05	0.094
Over sixty	30	0.072	0.066	0.10
Low income	13	0.097	0.06	0.15

(a) Descriptive statistics

	sum. sq.	df	F	p-value
group	0.007	2.0	1.21	0.298
Residual	0.55	221.0	NA	NA

(b) One way ANOVA

Table 7.5: Descriptive statistics and one way ANOVA for the Theil indices of activity transition matrices

ter is the same for nearly all outliers of the unconstrained and > 60 population group.

The ANOVA test for the distribution of individual mobility transition matrices (table 7.6) has rejected the null hypothesis ($F - value = 8.908, p = 0.0002$) which translates to the statement that the means of the Theil distributions are different. Looking at the mean Theil values for all three population groups in figure 7.26b, the > 60 and low income groups seem to have similar inequality levels ($\bar{T} = 0.21$ for low income, $\bar{T} = 0.20$ for > 60). However, for the individuals with Theil values belonging to the tails of the distribution, the levels of inequality seem to be particularly high. The qualitative profile of those individuals is similar to the above (mixture of employment statuses and residing in Outer London) with very high transition probabilities of using a particular mode (*Bus* or *Rail*).

group	**count**	**mean**	**std.**	**75 perc.**
Unconstrained	181	0.12	0.09	0.19
Over sixty	30	0.20	0.10	0.27
Low income	13	0.21	0.17	0.29

(a) Descriptive statistics

	sum. sq.	**df**	**F**	**p-value**
group	0.17	2.0	8.908	0.0002
Residual	2.22	221.0	NA	NA

(b) One way ANOVA

Table 7.6: Descriptive statistics and one way ANOVA for the Theil indices of transportation mode transition matrices

7.4 Discussion and conclusions

As already noted in section 7.3.1, the link between social exclusion and transport disadvantage is a complex one that has been approached through different ways in the literature. Having as a starting point the results of the CAA model, the case study using the Oyster/LTDS data showed that there are explicit links between limited access to opportunities and reduced access to public transport which can be further explored using passive unlabelled mobility data.

The general trend for distribution of activities throughout the day is similar in the three population groups, a fact which is not surprising given the fact that activity types were imputed, and not observed, from secondary data. However, including assumptions about the nature of the sociodemographic background of individuals allowed the shaping of head and tails of these distributions, particularly for *Employment* activity type.

Comparably, the low income population group was found to have smaller probabilities of activity type *Eating and Drinking* which could be linked to the reduction of the capability of using the public transport for reaching entertainment related activities. Although it is hard to make assertions due to the reduced sample size, it is nevertheless interesting to observe that the activity type that has the lowest probability range is the one that is the most elastic compared to *Employment* or *Education and Health* for example. On the other hand, the probabilities of activity type *Education and Health* are higher for the low income group. Considering that a large number of individuals in this population group are students, this fact comes as no surprise.

For all population groups, activity type *Outdoors and Recreation* was found to have the lowest probabilities compared to the rest of the activity types. Besides the limited number of POIs attributed to this category type, the RODS predicted probabilities using the Oyster card sociodemographic data placed on this category as a prior were weak relative to the rest of the activity types (the mean value of performing activity type *Outdoors and Recreation* was ≈ 0.13 for all individuals in the Oyster sample). This finding, together with the fact that *Outdoors and Recreation* had the highest accuracy of prediction (following *Employment* see section 6.4.1.1) makes the assertion of absence of such activity types from the dataset plausible, as opposed to being an artefact of the modelling process. This point is further discussed in section 8.4 where a discussion around the modelling and data limitations is presented.

A further breakdown of results can be made by examining the distributions of transportation mode use in relation to each activity type. For the unconstrained

population group, this revealed increased diversity in the probability distributions among individuals on the relative use of public transport, particularly for *Eating and Drinking* and *Education and Health* activity types. This could be an indication of the varying capability levels experienced by different people when reaching these activities. Overall, for this population group, the probability of using *Rail* services to reach the different activity types is higher compared to the rest of the modes. The exception is *Education and Health* where *Bus* seems to be higher ($\approx$ 0.42 for *Bus* and $\approx$ 0.36 for *Rail*). Assuming that these activities predominantly map to individuals who are either students/pupils or people that require medical care, decreased levels of *Rail* use compared to *Bus* provides evidence of reduced capability of using the *Rail* services by those individuals.

The use of *Bus* services is also the predominant mode of transport for the low income group for all activity types, a finding that is in line with existing evidence (Transport for London 2011). This should hardly come as a surprise, as cost is a significant barrier to transport in London, particularly for rail services. Other reasons mentioned in the literature for increased bus use is the possession of bus/rail cards for the low income groups, however, judging from the LTDS profile, none of the individuals in the group reported possessing one. The predominant use of bus for people in this group could also be the main reason for the observed geographical pattern, which is characterised by a tendency to avoid inner London (see figure 5.11). This fact, combined with reduced participation in activities as determined by the increased Theil index for this group, provides evidence of transport disadvantage compared to the rest of the groups.

Furthermore, the distribution of transportation mode use for the *Employment* activity type has been found to have a distinct multimodal shape throughout the day, characterised by relatively high probabilities in the morning and afternoon. This pattern could be explained by the mixture of employment statuses of the individuals in this group: full-time employed, part-time employed and students.

Compared to the rest of the groups, individuals in the over sixty sample have wider distributions of public transport use throughout the day, spanning a temporal window between morning and early evening. It is difficult to interpret this shape of transport use distribution, as from a data driven perspective, this group had the least number of transactions on average ($\approx$ 30 transactions per individual) compared to the rest of the groups ($\approx$ 32 for low income and $\approx$ 38 for the unconstrained group) which contributes to increased uncertainty of estimates. This fact coincides with evidence of non-travel (people who do not make trips) for this population group (Transport for London 2011) in London. From this perspective, it is difficult to make assertions of increased capability of using the public transport

network throughout the day compared to the unconstrained population group.

In terms of individual levels of equality as determined using the Theil index, the distribution of individual Theil values for the low income group is characterised by a statistically significant larger mean compared to the rest of the population groups. This provides evidence of increased levels of inequality experienced by this group, as the range of activities that are being reached is narrower. Examining the sociodemographic variables of the tails of this distribution (individuals belonging over the 75% of the Thiel distribution, a total of 10 individuals) using the matched Oyster card and LTDS IDs, further probing of the personal characteristics contributing to exclusion from activities and access to transportation modes can be deduced: 90% belong to black, Asian and minority ethnic backgrounds, 70% are women, all of them report income earned below £15,000 and all of them reside in outer Greater London. Moreover, the labour profile for these individuals is more unstable, with 8/10 people being either part-time employed or students.

In contrast, the sociodemographic profile of the Theil values of the tail of distribution for the unconstrained population group (16 individuals) consisted of 44% belonging to black, Asian and minority ethnic backgrounds, 62% women, all of them earning 25,000£ or more and 67% residing in outer Greater London areas with 15/16 being full time employed. The sociodemographic profile of the outliers of > 60 population group is similar to the unconstrained group (13 individuals) with 63% women, all of them earning 25,000£ or more and 80% residing in outer Greater London areas and 9/13 individuals being full time employed. However, the ethnic background of the individuals in this group is different with 10% belonging to black, Asian and minority ethnic backgrounds with the rest being White/British white. Judging from the above profiles, it is clear that the low income group is characterised by most of the risk factors that could result in social exclusion.

The general pattern of transition probabilities between activity types was found to be similar in the three population groups. One notable exception was the relatively low transition probabilities of *Employment* for the low income group. Empirically, this pattern could be attributed to the nature of working status of the individuals in the sample, half of whom were students, 25% full-time and 25% part-time employed. Moreover, this population group was found to have a larger distribution mean, providing evidence of less uniform transitions between activities. Indeed, examining the outliers of the Theil distribution, one notices increased transition probabilities for either education or employment related activities, depending on the individual's employment status.

Looking at the results of the mobility transition matrix, a number of interesting transportation habits are revealed. For a significant number of individuals

belonging in the unconstrained population group, the use of *Rail* services seem to persist throughout their trajectory, characterised by high *Rail/Rail* and low *Rail/Bus* transition probabilities. On the other hand, the inverse seems to be true for the > 60 and low income population groups, with high *Bus/Bus* and low *Bus/Rail* probabilities. This transition pattern is less uniform for the low income group, judging from the increased Theil values. Although it is difficult to make assumptions on the drivers behind this modal split pattern, it seems that, besides factors commonly mentioned in the literature such as egress and waiting time (Fearnley et al. 2018), sociodemographic factors (such as employment status and income) also play a role in the modal split patterns of individuals. In each case, the lack of sensitivity in switching between different public transport modes could be regarded as a reduced capability of using the public transportation network that can result in transport disadvantage.

The results from row-wise multinomial regression for modelling the mobility transition matrix showed that for many individuals in the unconstrained group, *PTAL* and *Trip duration* have a significant covariate effect. Furthermore, *Trip duration* coefficient was found statistically significant and positive for the majority of individuals for the > 60 and low income population group. This provides evidence stating that increased travel time accounts for increased transition rates from bus to all other modes. The effect of the *PTAL* covariate on transition sequences was found to be marginally significant and positive for the majority of individuals of all population groups. Contrary to the unconstrained group, the magnitude of this coefficient was larger for bus related transitions for the low income and > 60 groups.

The effect of chosen covariates for modelling the activity transition sequences was less profound, making assertions related to a population level contribution of external factors difficult. The individual based modelling approach however, allows identification of people that might experience a relative disadvantage, depending on the sign and magnitude of coefficients such as *IMD*. For example, in both the unconstrained and over sixty population groups, there is a considerable number of individuals with significant positive coefficients for the covariates related to transitions from *Retail* and from *Employment*. From a data driven perspective, this result makes sense since the majority of imputed activity types were related to *Retail* and *Employment* transition sequences. Empirically, this could also be interpreted as a relative disadvantage considering that these individuals are accessing activity types that are consistently located in areas characterised by high deprivation.

Travel diary driven assumptions about activity type inference have been found

to follow the passive mobility data sufficiently, judging from the overall alignment between prior/posterior for the majority of individuals in the unconstrained population group. However, disaggregation using age and income revealed deviations from the overall prior shape, which were larger for the *Employment* activity type. Lack of prior/posterior alignment was more profound in the case of potential mobility inference for the over sixty and low income population group. For the former, the probability of using the *Bus* services was increased after observing the data while for the latter the probabilities of different transport use was reshuffled in a way that characteristic public transport use signatures emerged for each individual user, providing further evidence of the need for a more granular approach on transportation mode modelling. To this end, the use of passive mobility data appears to be a promising alternative.

7.5 Chapter summary

In this chapter, the results of the CAA model introduced in section 6.4 were presented and evaluated in the context of exploring the link between social exclusion and transport disadvantage. For this task, unlabelled data from London's Oyster card dataset were used, linked with sociodemographic characteristics from London's LTDS. Building on the inferential capabilities of Bayesian networks, a set of latent nodes were defined and used to represent an individual's interaction with activity types and public transport modes given their sociodemographic profile and characteristics of the built environment. Results were compared and contrasted between three population groups: low income individuals, people over sixty years old and an unconstrained population group using Theil's index.

The results show that applying the proposed methodological framework to the Oyster card dataset can reveal distinct accessibility/mobility patterns at an individual level, in spite of the limited temporal window of observations. Using the dataset's LTDS association, these patterns can be related to individuals with a disadvantaged sociodemographic profile, providing further evidence on the link between social exclusion and transport disadvantage.

Chapter 8

Conclusions

8.1 Chapter overview

This chapter presents the conclusions of this thesis. In section 8.2 a summary of each chapter is provided, focusing on the main outcomes. The following section (8.3) provides the potential contributions to the literature across the different fields that this thesis touched. Section 8.4 discusses the limitations of the proposed approach, both from a modelling and from a data perspective. Finally, section 8.5 presents a general conclusion and outlook.

8.2 Thesis summary

The introduction of this thesis (chapter 1) defines the scope of this thesis. Several challenges have been identified related to both the conceptual framework on which individual accessibility measures are based and on the challenges of using unlabelled mobility data to apply such measures in empirical studies. Regarding the former, the main challenge relates to the difficulties in expressing individual accessibility within a modelling framework that can account for the complexity of the interacting factors that shape the individual accessibility components, namely using different transportation modes and reaching different destinations-opportunities. The latter refers to one additional modelling requirement: the extraction of high level semantic information related to accessibility components from low level mobility data. This is of particular importance given the existence of individual trajectory data from different sources and, at the same time, a considerable challenge given the low spatial and temporal resolution and the noisy nature of such data.

Chapter 2 provides the research background of this thesis. In particular, it elaborates on the role of accessibility in issues such as transport related social ex-

clusion and the link of accessibility to the different theories of social justice. Moreover, a description of the most commonly used numerical accessibility measures was given together with their strengths and weaknesses, along with a discussion of how these were applied in the context of assessing issues such as social exclusion and transport disadvantage, with special focus on the use of unlabelled mobility data. The chapter concluded by postulating that expressing accessibility through the lens of the capabilities approach (CA) is a promising alternative for this task.

This assertion was further explored in chapter 3. In this chapter, a brief overview of the CA is provided, together with a literature review on the application of this approach within the wider transportation literature. The common theme of the reviewed studies is that the CA does provide a theoretical framework upon which the different components and the defining factors of accessibility can be expressed. Furthermore, in this chapter, the potential of using existing numerical accessibility measures within the CA was discussed. It was argued that existing individual accessibility measures fail to capture the causal structure and interactions of the personal characteristics/external environment/capabilities/functionings relationship. In response to this, the chapter concluded by proposing the use of probabilistic graphical models for this task.

Chapter 4 started by providing some fundamentals of graphical models together with a basic categorisation of the different types of models. By examining their suitability in terms of expressing accessibility through the causal structure dictated by the CA from unlabelled mobility data, dynamic Bayesian networks were selected as the most promising approach.

Chapter 5 introduces the datasets used in the case studies of this thesis along with the required preprocessing steps. These included mobility datasets from a bespoke developed smartphone app, mobility data from the social media application Foursquare and mobility data from London's smart card system. For the latter in particular, preprocessing steps included constructing a hybrid dataset within which Oyster card transactions with bus boarding information. The resulting dataset is combined with socio-demographic LTDS data to produce individual trajectories with background personal characteristics information. Furthermore, this chapter introduced the algorithm used for inferring bus and tram alighting points by means of trip-chaining. Moreover, the notion of activity spaces was introduced as the basic unit for activity type inference along with the weighting function that was used to differentiate between *Employment/Education* and the rest of the activity types.

Chapter 6 presents the overall methodology for the formulation of the Capabilities Approach to Accessibility (CAA) model. In particular, section 6.2 intro-

duces the base dynamic hierarchical Bayesian network framework used in CAA, in the context of transportation mode detection using data from low resolution smartphones combined with travel survey data. Within this specification, information related to individual personal characteristics is included at the top level of the hierarchy as a prior distribution, under the assumption that such characteristics persists throughout an individual's trajectory. External covariates, on the other hand, are modelled through the rows of a transportation mode state transition matrix following the assumption that such space dependent variables change throughout an individual's trajectory. It was found that, depending on the individual's mobility characteristics, assumptions on the degree of influence of personal characteristics and external covariates based at population level secondary data do not always hold. Moreover, the transportation mode transition dynamics for each individual reflected on the probabilities of each row of the transition matrix revealed some insights on the extent of use of public transport by each individual. In terms of transportation mode detection accuracy, the proposed model was found to outperform the most commonly used classifiers for this task: RF, SVM and ANN (MLP). This section is based on the findings of Bantis & Haworth (2017).

In section 6.3 the proposed dynamic Bayesian network was reformulated in the context of activity type inference using Foursquare check-in mobility trajectories and POI data. Here, the main focus was the performance of the model in terms of activity type inference under different isochrone configurations. The model was found to achieve accuracy which is on a par with the reported accuracy thresholds in the literature for the 5 and 10 minute isochrone levels, in spite of belonging to the unsupervised classification models family. However, this performance is highly dependent on the nature of activity types. In particular, activity types associated with sparse POI vectors or class confounding POIs seem to produce results that are indistinguishable from a random classifier, or in some cases introduce systematic classification errors. In such cases, the degree of prior belief as expressed through the prior distribution plays an important role, as it tends to dominate over the sparse data likelihood. The results of this chapter are published in Bantis & Haworth (2019). Section 6.4 consolidates the results of sections 6.2 and 6.3 and introduces the CAA model. All components of the model are explicitly defined (capabilities/functionings/personal and external characteristics) along with the input/output of the mode. Various elements appearing in this section have been published in Bantis & Haworth (2020).

In an attempt to investigate how the notions of capabilities and functionings have been approached within transport disadvantage, chapter 7 provides the in-

ference results of CAA model using passive mobility data (Oyster card) and travel survey data (LTDS/RODS) for each node of the model. In section 7.2 The posterior quantities for all model's nodes are disaggregated by individual/population group. In general, the overall trend of potential accessibility (as expressed through the posterior distribution of activity types for each individual) appears to be similar. Closer examination, however, allowed some key differences to emerge between the population groups. In particular, the posterior probabilities of the low income group appear to be lower for "elastic" activity types, such as *Eating and Drinking*. Moreover, activity type posterior results demonstrated that, depending on the individual, there exists considerable variability in the probabilities of activity types throughout the day, even within the same population groups. Disaggregating the results further to account for the transportation mode used, the differences are more pronounced. Again, within-group variations are significant, however overall, use of *Bus* services is the predominant mode of transport for the low income group, as opposed to *Rail* services for both the unconstrained and the over sixty group. Furthermore, the transition probabilities for the low income group revealed a low transition rate between *Employment* and the rest of the activities, reflecting the employment status split of the individuals of this group (mix of students, part-time and full-time employed). The probabilities for transportation modes were higher for *Bus* related transitions for both the low income and over sixty population group. The effect of the chosen external variables for both activity type and transportation mode transitions were found to vary across individuals considerably. Finally, the degree of prior belief on the effect of personal characteristics was consistent with their posterior counterparts for activity types (with the exception of *Outdoors and Recreation* activity). However, for transportation modes there was a considerable deviation for the low income and > 60 group between the prior/posterior correspondence, essentially reshaping the logistic regression derived assumptions related to the choice of transportation mode from sociodemographic characteristics.

Finally, section 7.3 provided a framework for evaluation of transport related social exclusion using the model's posterior quantities over the elements of the capabilities set based on Theil's index. Individual indices per population group (low income, over sixty years old and unconstrained group) were benchmarked against maximum entropy, which translates to the theoretical conditions of complete equality. The sociodeomographic profile of individuals with the highest inequality levels correlated with those of low income, non white ethnic backgrounds residing in outer Greater London areas for nearly all elements of the capabilities set, suggesting that these individuals could experience higher risk of transport

related social exclusion.

8.3 Contributions to the literature

This section highlights the potential contributions to the literature from the outcomes of this thesis. These are divided in two areas: contributions to accessibility literature and contributions to transport literature.

8.3.1 Contributions to accessibility literature

The formulation of accessibility through a CA framework potentially addresses the following challenges in accessibility literature:

1. A data driven/graphical model approach to CA operationalisation was introduced.
2. By using the output of the model, justice theoretic evaluations based on equality can be performed at an individual level.
3. Combining accessibility and mobility, a framework for multimodal accessibility per activity type categories is introduced.
4. Using individual level trajectory data, the proposed model allows re-evaluation of prior assumptions on travel behaviour based on empirical evidence.
5. Using unlabelled mobility data of low resolution, particularly from transportation network providers, the scope of accessibility appraisals is broadened.

8.3.2 Contributions to transportation literature

Through the use of low resolution unlabelled data the proposed modelling framework contributes to addressing the following challenges in transportation literature:

1. By including personal characteristics and variables related to the wider environment, improved transportation mode detection accuracy compared to commonly used classifiers is achieved, even with a limited observation feature vector.
2. By using a hierarchical dynamic Bayesian network over POIs bounded by an isochrone in an unsupervised classification setting, activity type detection accuracy can be achieved that is on a par with supervised classification

methods. This is of significant importance considering that ground truth information is absent for most service provider generated mobility data.

3. By using ground truth Foursquare data, the limits of activity type detection accuracy was assessed for data of increasingly lower resolutions (such as AFC, cell-tower data). This is done by evaluating the model's performance at different isochrone levels.

4. By modelling the transitions between transportation modes/activity types using row-wise multinomial regressions, the proposed model allows the explicit inclusion of external variables in the evolution of state space.

5. By modelling the Dirichlet distribution concentration parameters using external data, personal information related to mobility/accessibility patterns can be introduced in the modelling process.

8.4 Limitations

In this section, some limitations of this thesis are identified and elaborated focusing on two areas: modelling framework and data limitations. These could serve as opportunities for further development of the proposed methodology.

8.4.1 Modelling limitations

Although considerations regarding the spatial and temporal structure of observations were implicitly included in the modelling framework through the use of space (eg. isochrones at an alighting point and space varying covariates) and time dependent mobility data (eg. Oyster card, geo-location data), explicit spatial and temporal correlations between observations have not been accounted for. In the discrete case of public transport access points as areal units, these correlations could be caused by the spatial configuration of bus stops/rail stations, and have been found to be statistically significant at an aggregated level (Bantis et al. 2015). Given the nature of the model where the focus is on the individual, one option would be to include a spatial autologistic term in the row-wise multinomial regressions of transition matrices. During the experimentation phase of this study, such terms were introduced by imposing a neighbouring structure defined by a 5km threshold at each access point and a 2 hour temporal window between observations. However, this introduced added complexity in the model making MCMC sampling extremely long and convergence difficult. Furthermore, due to the limited number of unique destinations per individual trajectory, information

sharing between neighbouring units was found to be weak (as evidenced by the non-significant spatial and temporal posterior effect). As a result, it was decided not to include such terms in the final model.

Another modelling limitation that should be taken into account when interpreting results particularly for the posterior distributions of activity types, is related to the nature of unsupervised classification framework adopted in this study. Within this, activity type inference is solely determined by the distribution of POIs within an isochrone, the prior assumptions on the individual's propensity to perform an activity type given personal characteristics and characteristics of the external environment. For activity type *Outdoors and Recreation* all these data sources fall back compared to the rest of activity types. As a result, for nearly all participants, the posterior probabilities of performing this activity type fluctuated around the random threshold of *0.2*. If the focus is making assertions regarding this particular activity type, extra considerations need to be taken in the calculation of the likelihood. As an example, this could be expressing an extra source of information, such as being weekday or weekend. Another way is to weight the observation vector by a POI capacity variable, using social media data such as Twitter. This could be achieved by modifying the informative prior on the activity type probabilities. If such a modification is not possible, then an option is to include an extra likelihood term (eg. through soft data-factor potential) increasing the likelihood of this activity type according to the capacity of POIs and/or type of day (weekend/weekdays) and decrease it otherwise (Jordan et al. 2004).

8.4.2 Data limitations

In all three case studies in this thesis, passively generated mobility data of varying spatiotemporal resolution were used to infer accessibility/mobility quantities and readjust the prior assumptions in light of evidence. As such, any generalisations to population level characteristics should be made keeping this important consideration in mind. This is especially true for the case study using Oyster card data where transport related equality claims were made between three population groups: a low-income, over sixty and the unconstrained population group. As the sample size of these groups is unequal, it is important that considerations related to homogeneity of variance hold, particularly for the distribution of activity types where the one-way ANOVA test turned out to reject the null hypothesis (populations with equal mean). Repeating the test with a non parametric version of ANOVA (Kruskal-Wallis H-test) has also resulted in rejection of the null hypothesis (statistic=7.48, p-value=0.023) increasing the confidence of the equality assertion made in section 7.3.

Furthermore, the varying sample size of Oyster card data for the different population groups has an impact on the geographic representativeness of the study area. While for the unconstrained and over sixty sample the visited locations appear to be uniform throughout the study area, the visited locations of the low income group appear to cluster radially across the study area (see Figure 5.11). Although a positive correlation exists between the index of deprivation and the visiting locations of the low income group (OLS slope 0.013 compared to a negative correlation for the unconstrained group with OLS slope -0.044 and the over sixty group with OLS slope -0.001) which intuitively is what would one expect, it is difficult to make any firm assertions regarding the geographic representativeness of this population group.

With regards to the temporal extent of chapter 7, the available dataset did not allow any deeper evaluation of the way individuals adjust their activity/travel behaviour in the face of an event that could impact accessibility. Such an event can be related to personal characteristics (such as a change in employment status) or can be infrastructure related (for example, an introduction of a new public transport connection. A future direction could involve using an extended time span together with information on significant events to assess whether an adaptation of behaviour is represented in the evolution of mobility/accessibility nodes of the model.

Moreover, the absence of tram journeys from the sample can be considered a direct result of the state of tram services in the city, where tram journeys represent less than 1% of the total journeys made each year (Transport for London 2019). Considering that the final Oyster card/LTDS sample of 224 individuals was around 2.4% of the total Oyster card sample provided, the chances of encountering individuals using tram services as the primary mode of transport are slim. As a result, to be able to make claims on mobility/accessibility patterns of tram users, stratified sampling is needed targeting this subgroup explicitly.

Finally, it should be noted that the results of chapter 7 are bounded by the quality of information provided by the Oyster/LTDS sample. For example, population groups that are thought to present high risk of transport related social exclusion such as the unemployed, disabled and retired were not represented in the sample. It would be of great value if the analysis was repeated with these groups, as it would demonstrate the degree of robustness of the proposed methodological framework.

Related to the above, it is important to note that while chapter 7 represented an individual's potential accessibility to activities and potential mobility using public transport through the Bayesian network structure, access to an individual's actual

"wants" and "desires" behind their choices remains out of reach and can only be uncovered through extensive qualitative studies. For example, people with lower income may choose to eat and drink out less due to lack of resources. Transport accessibility may be a factor, but its effect might be exaggerated by the lack of access to the drivers behind the choices made by those individuals. Nevertheless, the current structure of the proposed model could be used to identify deviations from the average equality levels so that further investigation can be undertaken. Future directions will be steered towards a qualitative validation of the findings.

8.5 Conclusion and outlook

In this thesis, a novel approach to evaluating individual accessibility was proposed by framing the modelling methodology through the CA. Following the hierarchical structure of the CA, the different components that shape an individual's ability to reach opportunities were explicitly modelled in a probabilistic way through the notions of latent capabilities and observed functionings. The potential of the proposed methodological framework to evaluate individual based transport based social exclusion was assessed through a case study using London's AFC data. It was found that the proposed framework could identify individuals that exhibit high risk of social exclusion by comparing the distributions of the capabilities sets.

The implementation methodology was based on dynamic Bayesian networks using low resolution mobility data from different sources. The hierarchical nature of the model allows the incorporation of assumptions related to mobility/accessibility behaviour while at the same time the dynamic nature of the model was used to include characteristics of the built environment that change throughout an individual's trajectory. The model has been applied to data of different spatial and temporal resolution and exhibits strong performance in both transportation mode and activity type inference compared to existing models.

Abstraction of the proposed model to the entirety of service provider's data is challenging but achievable. It is challenging due to the level of detail of the model, together with the nature of probabilistic inference. Achievable due to recent advances in sampling schemes (e.g. Hamiltonian Monte Carlo, Variational Inference) that have the potential of reducing sampling autocorrelation and achieve faster convergence.

Further to the potential of transportation service provider's data, London and other cities across the world are encouraging transactions using contactless payments instead of AFC data. This fact, together with the wide adoption of contactless payments for consumer purposes, would enable merging of mobility patterns

with consumer habits, providing insights to an often overlooked element of accessibility which is access to non-employment recreation/retail/consumer activities.

The use of low resolution data in the case studies presented in this thesis enables the application of the model to datasets of increased population coverage such as data obtained from mobile phone network operators. Indeed, the isochrone approach adopted in the derivation of the activity type observation vectors mimics the spatial resolution of GSM cell-tower data obtained in urban areas. Furthermore, the use of such data would allow for a more refined definition of the mobility component of accessibility that goes beyond the use of public transport. Questions such as the following could then be interrogated: To what extent active transportation (e.g. bicycles) compliments more traditional modes of transport? Do they expand the range of activity types reached? What is the role of mobility as a service applications (MaaS) such as Uber in the ability of individuals to reach activities, particularly for vulnerable population groups such as the elderly? Interrogating these questions would require reformulation of the model within the context of transportation mode detection described in section 6.2.

Related to the use of the potential data sources described above, it is important that any modelling attempts are not compromising aspects of an individual's right of personal information. This is especially true for data of increased fidelity such as mobile-phone operator data. Careful consideration needs to be taken to prevent any prospects of individual identification through their trajectories. Appropriate aggregation should be applied, ideally within pre-defined census boundaries. Such a step would provide the additional benefits of directly relating the individual trajectories to census data, as well as allowing the incorporation of spatial statistics methods such as accounting for spatial autocorrelation between the census tracts.

Furthermore, although the case studies presented use London as the city of reference, the proposed modelling framework could be applied to other cities in the UK or internationally, provided that data on mobility trajectories and activity type proxies exist. To this end, the ubiquitous use of smartphones by the majority of population makes the application of the model to other cities plausible. In this way, accessibility evaluations across different cities can be achieved, providing the background for assessing transport related social exclusion across different cities.

Finally, it is worth pointing out that the proposed model has the potential to be used in areas beyond assessing individual accessibility, and into the realm of urban, transportation planning and behavioural modelling. As the output of the model is mobility and activity type patterns, these can be used to spot development opportunities as well as to provide an estimate of the capacity of utility services required (such as electricity, connectivity etc.) for those. City planners

and developers can then assess the potential to invest in new businesses and homes in those areas, depending on the levels of access and activity participation. Within transportation planning, the output of the model can directly inform about the need to expand or complement the existing transportation network, depending on the demand of available activities in a given area. Within the field of behavioural modelling, the CAA model can provide the input needed for creating realistic agents within a simulation context (e.g. modelling the responses of people during a pandemic). This is achieved through the explicit causal link between passive mobility data and sociodemographic characteristics, which allows the extraction of detailed dynamic spatiotemporal activity and transportation signatures from unlabelled data.

In light of the ever increasing trend of urbanisation, accessibility is likely to be a major problem for future cities, as current infrastructure will be stressed to accommodate the needs of an increasing urban population. With the levels of inequality in transport likely to increase as a result of competition for resources, policy makers will need more information on the causes of transport related social exclusion. To that extent, new technologies combined with big data that provide interpretable results could provide evidence to promote equity.

Policy makers have a huge responsibility for promoting people's well-being in a fair fashion, especially in these politically unstable and uncertain times. Decision making is a difficult task, and in that line of argument, policy makers will need to move away from aggregated indices and embrace the complexity of transport related social exclusion as a phenomenon, in order to promote a fairer public transport system for everyone. Accessibility as a concept has the potential to make significant contributions to this, and as such, it should continue to evolve both in theoretical formulations and implementation methods.

Appendix A

Results of Logistic regression for LTDS

The intercepts were 15.81, 3.01, -6.87 for the categories *Bus, Rail, Tram* respectively.

	Bus		**Rail**		**Tram**	
Variables	e^{β}	p-value	e^{β}	p-value	e^{β}	p-value
Disability Type: Wheelchair user	0.441	0.0006	0.028	0.5762	1.274	0.5689
Disability Type: Mobility	0.747	0.0108	0.000	0.6314	0.759	0.2205
Disability Type: Visual	0.851	0.6946	0.430	0.7942	0.000	0.9842
Disability Type: Hearing	1.032	0.9528	0.896	0.9535	0.000	0.9881
Disability Type: Learning	0.655	0.5019	0.128	0.5258	1.504	0.5367
Disability Type: Mental health (0607 onwards)	0.497	0.0211	0.017	0.5498	0.726	0.5644
Disability Type: Serious long-term (0607 onwards)	1.011	0.9610	0.389	0.8623	0.108	0.0282
Disability Type: Other	0.809	0.4822	0.070	0.6386	0.254	0.1773
Age	1.000	0.9538	0.000	0.9788	0.994	0.2214
Income: £100,000 or more	0.519	0.0000	0.000	2.0623	0.041	0.0019
Income: £15,000 - £19,999	0.920	0.5618	0.822	0.9755	0.959	0.8748
Income: £20,000 - £24,999	0.936	0.6395	0.674	1.0479	1.292	0.3112
Income: £25,000 - £34,999	0.843	0.1798	0.114	1.1752	1.268	0.3184
Income: £35,000 - £49,999	0.792	0.0615	0.081	1.1960	0.766	0.3126
Income: £5,000 - £9,999	0.927	0.5754	0.156	0.8686	1.004	0.9874
Income: £50,000 - £74,999	0.810	0.0904	0.000	1.5977	1.033	0.8979
Income: £75,000 - £99,999	0.594	0.0003	0.001	1.5509	0.304	0.0094

Income: Do not know	0.968	0.8003	0.002	1.3469	0.894	0.6301
Income: less than £5,000	0.923	0.6348	0.572	0.9359	1.057	0.8399
Income: Refused	0.740	0.0105	0.012	1.2666	1.165	0.4893
Sex	0.990	0.8569	0.000	1.4553	1.012	0.9177
Working Status: Part-time paid employment (less than 30 hours a week)	1.164	0.1092	0.192	0.8984	1.662	0.0150
Working Status: Full-time self-employment (30+ hours a week)	1.236	0.0351	0.000	1.4954	1.435	0.1712
Working Status: Part-time self-employment (less than 30 hours a week)	1.178	0.2657	0.080	1.2690	1.350	0.4167
Working Status: Student/school pupil	2.092	0.0000	0.477	1.0842	0.889	0.6712
Working Status: Waiting to take up a job	0.947	0.8703	0.369	1.3166	1.233	0.7811
Working Status: Unemployed and looking for job	1.704	0.0015	0.001	0.6610	1.091	0.7835
Working Status: Unable to work because of long-term illness or disability	1.924	0.0041	0.000	0.4652	1.377	0.4382
Working Status: Retired	1.268	0.0795	0.000	0.6052	1.055	0.8549
Working Status: Regular unpaid Voluntary Work	1.249	0.4222	0.721	0.9202	1.690	0.2584
Working Status: Looking after home or family	1.253	0.0993	0.000	0.4760	0.850	0.5938
Working Status: Other non-working	0.845	0.6844	0.647	1.1793	0.614	0.6418
Occupation: Middle or junior managers	1.094	0.5407	0.461	1.1055	0.801	0.5526
Occupation: Modern professional occupations	0.942	0.5598	0.085	1.1757	0.544	0.0200
Occupation: Routine manual and service occupations	1.436	0.0252	0.001	0.6649	0.666	0.2401
Occupation: Semi-routine manual and service occupations	1.248	0.1359	0.237	0.8672	0.805	0.4704

Occupation: Senior managers or administrators	0.902	0.3891	0.216	1.1534	1.328	0.3235
Occupation: Technical and craft occupations	0.783	0.0975	0.017	0.7249	0.804	0.5486
Occupation: Traditional professional occupations	0.829	0.1413	0.005	1.4350	1.338	0.3362
Household Members	0.971	0.1613	0.000	0.8336	1.136	0.0022
Has a driving license	0.845	0.1031	0.000	1.6414	0.684	0.0434
Free travel pass	0.533	0.0000	0.000	0.7073	0.564	0.0061
Bus pass: Monthly	0.627	0.9994	0.356	0.5795	0.545	0.9924
Bus pass: Not asked	0.000	0.9793	0.296	1.7793	0.2	0.9927
Bus pass: Weekly	0.616	0.9994	0.390	0.6106	0.2	0.9928
Ethnic group: Asian or Asian British - Chinese	0.567	0.0735	0.246	1.3747	13.771	0.0147
Ethnic group: Asian or Asian British - Indian	0.766	0.2958	0.136	0.7471	9.111	0.0312
Ethnic group: Asian or Asian British - Other Asian background	0.948	0.8435	0.634	1.1014	8.209	0.0417
Ethnic group: Asian or Asian British - Pakistani	0.794	0.4429	0.520	0.8600	2.920	0.3570
Ethnic group: Black or Black British - African	1.593	0.0903	0.844	0.9624	8.599	0.0359
Ethnic group: Black or Black British - Caribbean	1.349	0.2878	0.016	0.6140	12.250	0.0148
Ethnic group: Black or Black British - Other Black background	1.390	0.4073	0.857	0.9512	10.767	0.0298
Ethnic group: Mixed or multiple ethnic groups - White and Black Caribbean	1.157	0.7103	0.789	1.0820	6.662	0.1048
Ethnic group: Mixed or multiple ethnic groups - White and Asian	0.793	0.6113	0.144	0.5835	4.792	0.2745
Ethnic group: Mixed or multiple ethnic groups - White and Black African	0.682	0.4413	0.290	0.6796	33.272	0.0015

Ethnic group: Other Ethnic group - Any Other	1.249	0.5008	0.345	0.7947	0.000	0.9778
Ethnic group: Other Ethnic Group - Arab	4.577	0.0542	0.087	1.8129	6.152	0.1446
Ethnic group: Other Mixed or multiple ethnic background	1.948	0.1555	0.358	0.7397	6.268	0.1393
Ethnic group: Other White	0.833	0.4584	0.609	0.9090	3.748	0.1994
Ethnic group: Refused	1.151	0.8662	0.949	0.9646	0.000	0.9930
Ethnic group: White - English/Welsh/Scottish/Northern Irish	0.648	0.0696	0.001	0.5499	7.710	0.0437
Ethnic group: White - Irish	0.806	0.4673	0.171	0.7266	6.870	0.0752
Ethnic group: White - Other British	0.688	0.2293	0.430	1.2338	2.724	0.4182
Car use: 2 days a week	0.932	0.5506	0.905	1.0145	1.273	0.4320
Car use: 3 or 4 days a week	0.706	0.0027	0.277	0.8805	0.914	0.7718
Car use: 5 or more days a week	0.275	0.0000	0.000	0.4551	0.728	0.2749
Car use: At least once a fortnight	1.264	0.2194	0.018	1.6114	0.608	0.4366
Car use: At least once a month	1.949	0.0015	0.810	0.9554	0.742	0.6004
Car use: At least once a year	1.822	0.0003	0.124	1.2715	0.444	0.1178
Car use: Never used	2.206	0.0000	0.033	1.2901	1.012	0.9678
Car use: Not used in last 12 months	2.211	0.0000	0.783	1.0339	1.291	0.3985
Car as a passenger use: 2 days a week	0.898	0.2252	0.188	0.9027	0.736	0.1171
Car as a passenger use: 3 or 4 days a week	0.861	0.1450	0.432	0.9330	0.837	0.3979
Car as a passenger use: 5 or more days a week	0.865	0.2403	0.001	0.7135	0.826	0.4453
Car as a passenger use: At least once a fortnight	1.194	0.0988	0.105	1.1608	0.623	0.0517
Car as a passenger use: At least once a month	1.092	0.3310	0.913	0.9914	0.721	0.0984
Car as a passenger use: At least once a year	0.923	0.3291	0.892	0.9900	0.933	0.6899

Car as a passenger use: Never used	1.275	0.0692	0.155	1.1685	0.532	0.0553
Car as a passenger use: Not used in last 12 months	1.019	0.8636	0.396	0.9233	1.001	0.9978
Regular taxi use: 2 days a week	1.022	0.9366	0.808	1.0711	0.682	0.5940
Regular taxi use: 3 or 4 days a week	0.856	0.6388	0.285	1.5356	2.271	0.2251
Regular taxi use: 5 or more days a week	0.538	0.1751	0.154	0.5002	0.000	0.9889
Regular taxi use: At least once a fortnight	0.993	0.9727	0.408	1.2073	0.677	0.4735
Regular taxi use: At least once a month	0.981	0.9184	0.811	1.0480	0.575	0.2531
Regular taxi use: At least once a year	0.781	0.1641	0.039	0.6855	1.041	0.9247
Regular taxi use: Never used	0.795	0.2077	0.000	0.3882	0.854	0.7094
Regular taxi use: Not used in last 12 months	0.785	0.1781	0.000	0.4021	0.792	0.5827
Private taxi use: 2 days a week	0.918	0.7290	0.320	0.8204	1.484	0.3443
Private taxi use: 3 or 4 days a week	0.883	0.7416	0.043	0.5627	0.000	0.9824
Private taxi use: 5 or more days a week	0.339	0.0279	0.661	0.8094	0.000	0.9901
Private taxi use: At least once a fortnight	1.155	0.4294	0.177	0.8185	1.233	0.5275
Private taxi use: At least once a month	1.005	0.9769	0.405	0.8979	0.974	0.9301
Private taxi use: At least once a year	0.752	0.0592	0.008	0.7272	0.851	0.5551
Private taxi use: Never used	0.662	0.0104	0.016	0.7378	0.897	0.7112
Private taxi use: Not used in last 12 months	0.620	0.0022	0.000	0.5963	1.178	0.5602
Walking: 2 days a week	1.773	0.0006	0.232	1.2300	1.128	0.7779
Walking: 3 or 4 days a week	2.591	0.0000	0.000	1.7564	1.135	0.7458
Walking: 5 or more days a week	3.208	0.0000	0.000	2.6672	1.082	0.8232

Walking: At least once a fortnight	1.003	0.9923	0.276	0.7065	1.087	0.9168
Walking: At least once a month	0.406	0.0027	0.008	0.4209	1.264	0.7328
Walking: At least once a year	0.333	0.0006	0.740	0.8845	1.134	0.8562
Walking: Never used	1.326	0.5804	0.763	0.8270	0.000	0.9900
Walking: Not used in last 12 months	0.242	0.0000	0.104	0.5091	0.622	0.5059

Appendix B

Results of Multinomial regression for RODS

The baseline category was *work*. The intercepts were 0.00002, 0.0067, 0.0003, 0.000007 for the categories *Eating/Drinking,Education/Health, Retail, Sports/Entert.* respectively.

Variables	**Eating/Drinking**		**Education/Health**		**Retail**		**Outdoors/Recr.**	
	e^{β}	p-value	e^{β}	p-value	e^{β}	p-value	e^{β}	p-value
Time of day	2.4327	$<$ 2.2e-16	1.8640	8.841e-12	2.0164	1.731e-14	2.4212	$<$ 2.2e-16
Sex: Male	2.4094	0.5153	0.8637	0.8725	2.6512	0.4367	0.8794	0.9081
Sex: Female	2.2946	0.5378	1.3935	0.7136	4.1259	0.2561	1.0952	0.9345
Disability	2.4238	0.10190	1.7134	0.2611	1.9576	0.1685	3.9963	0.0135
Age	1.0117	0.9053	0.8184	0.0279	1.00068	0.9940	1.3403	0.00262

Appendix C

Posterior activity type distributions for all participants of section '6.3'

The purple line indicates ground truth check-in activities.

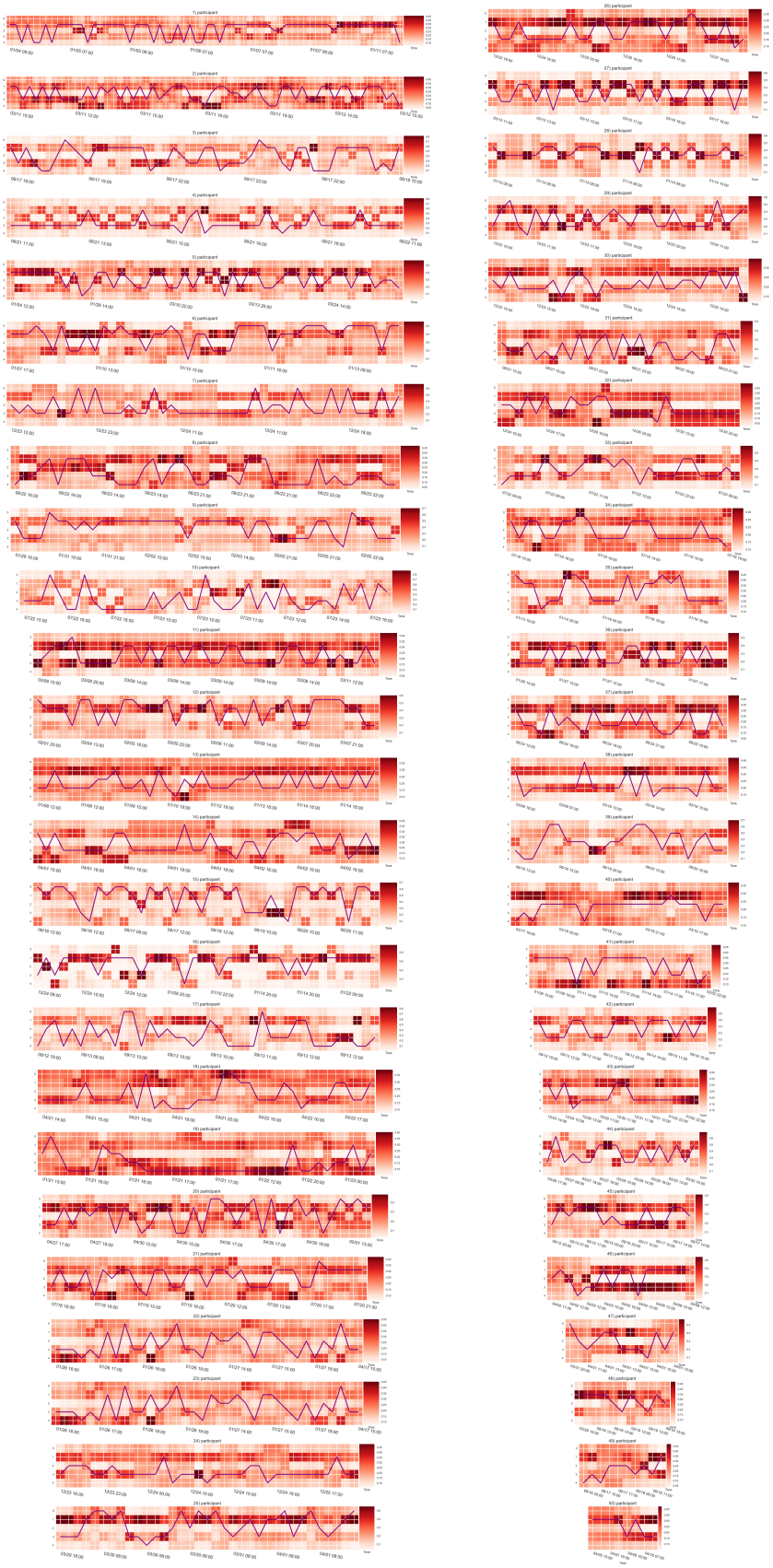


Figure C.1

Appendix D

CAA model convergence diagnostics

This section presents the convergence diagnostics for the MCMC simulations for all variables of figure 6.35. It begins by discussing model convergence using within and between MCMC chain diagnostics. It then provides an indication of the model's accuracy by using a limited self-labelled Oyster card dataset, before discussing the results of the simulations. The posterior quantities for the capability variables of section 6.4.2 are presented, grouped by the population groups specified in section 6.4.

D.1 Convergence diagnostics

Two methods of assessing the convergence of the MCMC simulations were considered: The first follows a within MCMC chain approach and uses visual inspection of the chains and the associated auto-correlation plots as well as Geweke's z-score, while the second adopts a between chain approach using the Gelman and Rubin $\hat{R}$ statistic.

D.1.1 Within chain convergence

The most straightforward approach of assessing MCMC chain convergence is by inspecting the chains for elements of asymptotic behaviour. If the sequence of samples appear to have a constant mean and variance, this could provide evidence of sampling from the target posterior distribution. This should be repeated for all stochastic variables as convergence of one does not necessarily imply convergence of the other variables. For this study, trace plots (plots of the simulated MCMC samples) as well as their auto-correlation plots are used as a means of qualitative

assessment of convergence (Lawson 2009). In line with the preceding case studies of this thesis, Geweke's z-score was used as a means of formal indication of within-chain convergence.

The remainder of this section presents the results for the > 60 population group. Results for the unconstrained and low income groups are presented in appendix E.

Inspecting the trace plots and auto-correlation plots for the 30 individuals of the > 60 population group for the d node of model 6.35. Figure D.1 below shows the results for all d_i^κ of each individual.

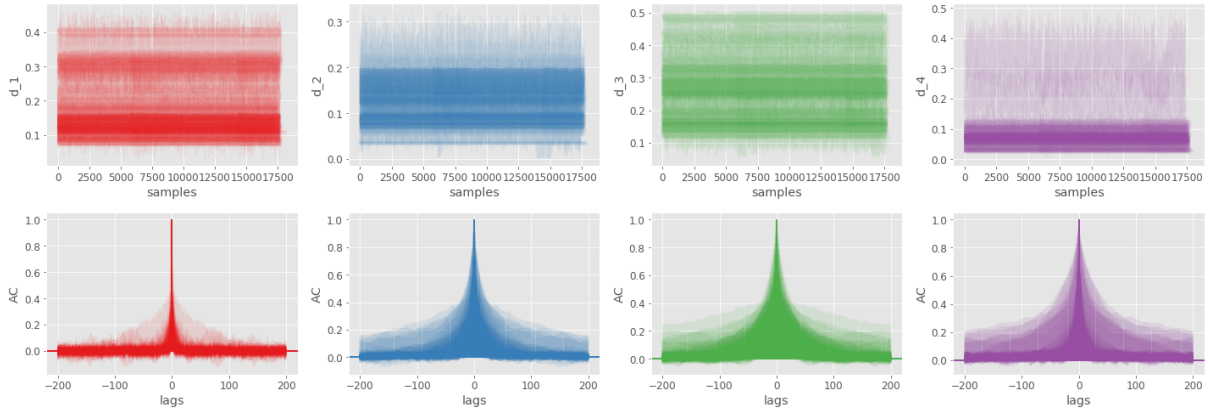


Figure D.1: Trace and auto-correlation plots of d_i^κ for all individuals in the > 60 group (Red: Eating and Drinking, Blue: Education and Health, Green: Retail, Purple: Sports and Entertainment. *2500 burn in applied*)

Judging from the trace plots, the MCMC chains do not seem to present any signs of non-stationarity. The auto-correlation plots suggest various degrees of correlation between subsequent samples, however this seems to decrease with increasing lags. High auto-correlation suggests that the MCMC algorithm explores the parameter space with reduced efficiency, which doesn't contribute significantly to the determination of the chain's statistical parameters. The process of "thinning" (keeping every other n sample of the chain) is often suggested to reduce auto-correlation in the chains (Kruschke 2010). However, some authors suggest that such remedy is only computational, reducing the effective number of samples used to approximate the posterior distribution (Brooks et al. 2011). In this study, a combination of different approaches was used to reduce auto-correlation, including thinning, different burn-ins and a combination of different MCMC algorithms for different nodes (eg. Metropolis-Hastings, Adaptive Metropolis).

Next, the Geweke's z-scores for all d_i^κ variables of the individual models were computed for the first 10% and last 50% of the chains. Figure D.2 below shows

the distributions of z-score values for all chains.

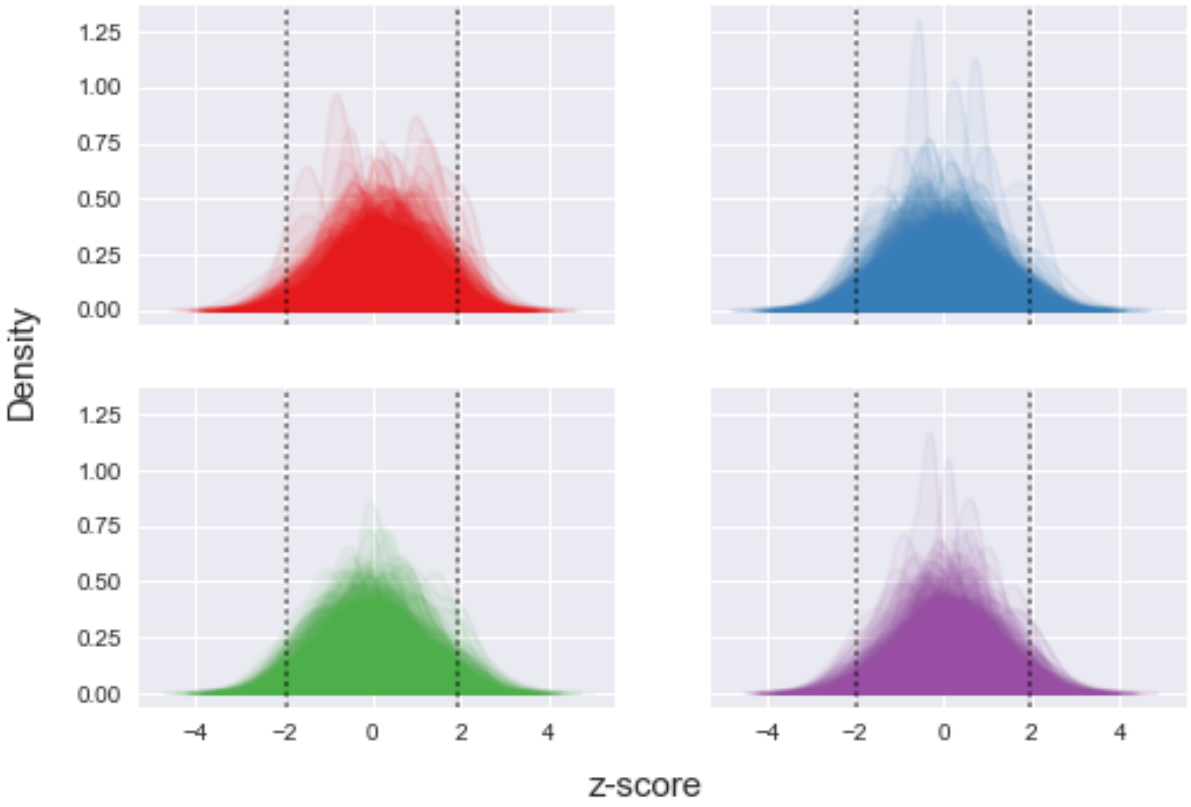


Figure D.2: Gewekes z-scores of d_i^{κ} for individuals in the > 60 group (Red: Eating and Drinking, Blue: Education and Health, Green: Retail, Purple: Sports and Entertainment)

As it can be seen, the bulk of the z-scores lie within two standard deviations of the mean, indicating convergence. For the remaining scores, we allow 5% of the calculated scores to lie outside this range due to type I errors of multiparameter significance tests. For the variables outside these ranges, Gewekes z-scores indicate that more samples are needed to provide an indication of chain convergence.

The same approach was applied for the m nodes of the model, corresponding to potential mobility of public transport means. Figures D.3 and D.4 below show the trace and Geweke's plots for *Bus* and *Rail* transport for all individuals over sixty years old in the sample.

The above plots present similar characteristics with d nodes, with the exception of some bus variables having relatively high auto-correlation (> 0.1). Reparametrisation of the models is suggested by some authors as a way to reduce auto-correlation (eg. variable zero mean centre) (Browne 2004), however, for this

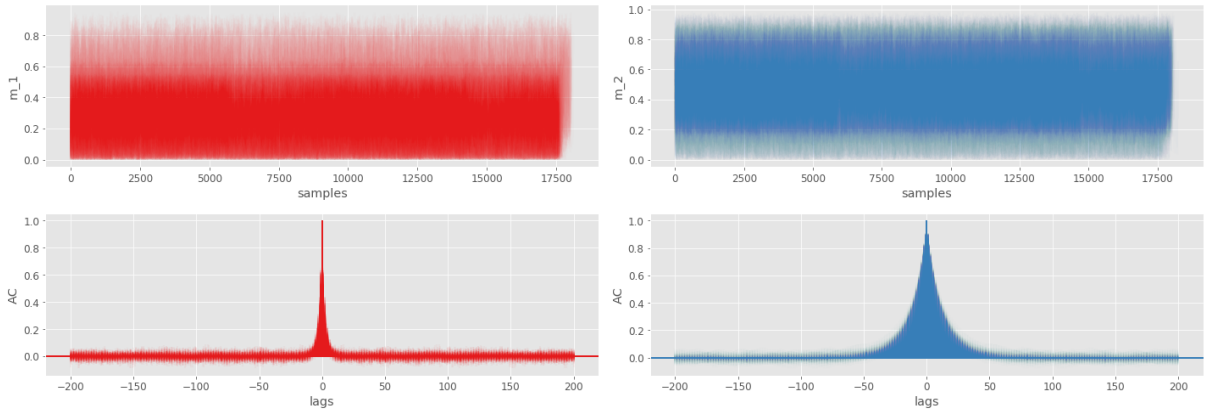


Figure D.3: Trace and auto-correlation plots of m_i^{κ} for all individuals > 60 years old in the sample (Red: Bus, Blue: Rail. *2500 burn in applied*)

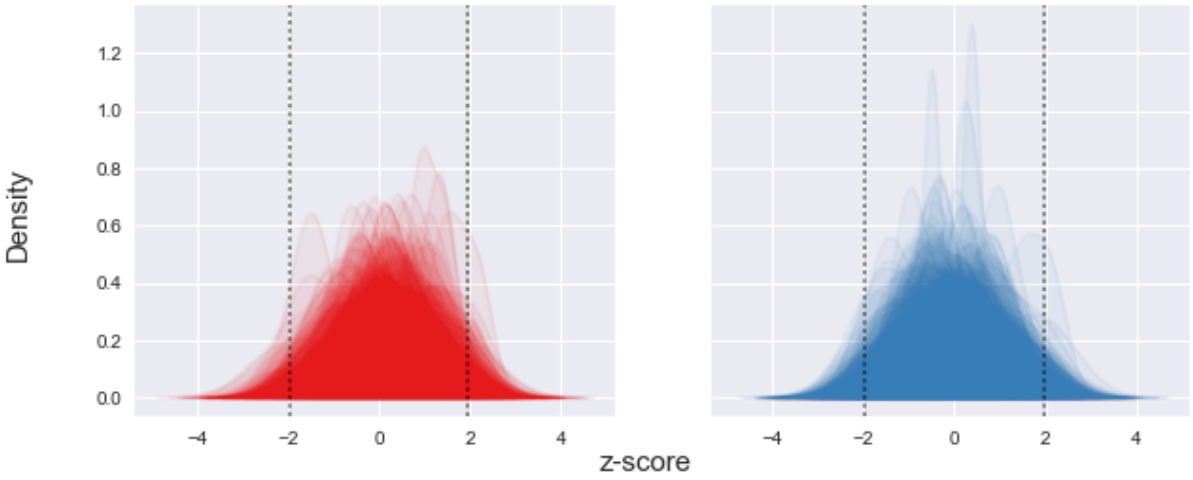


Figure D.4: Gewekes z-scores of m_i^{κ} for individuals over sixty (Red: Bus, Blue: Rail)

model is not an option due to the discrete nature of the mobility variables, so thinning and extending the length of the chains was applied as an alternative for reducing auto-correlation.

Next, the convergence diagnostics are provided for the transition matrices T_z and T_m (Figures D.5, D.6, D.7, D.8). Sampling from those nodes was more difficult as indicated the convergence plots. Since some of the external covariates used in modelling the sequence of transitions are correlated (particularly nodes such as trip duration and PTAL), they introduce collinearity in the β external parameters, making the MCMC simulations less efficient. In the context of Bayesian inference,

this translates to the fact that transition sequences don't provide enough information to explain the individual coefficients. In such cases, adjusting the scale of the MCMC algorithm depending on the acceptance rate can help towards more efficient sampling. On the modelling side, introducing stronger prior assumptions when forming the model can help differentiate the effect of the covariates and provide more stable estimates (Gelman et al. 2013). However, since there is no reliable source of information that could be used to introduce an informative prior for the covariates, a thinned version of the chain (keeping every other sample) was used to generate the coefficients' statistics.

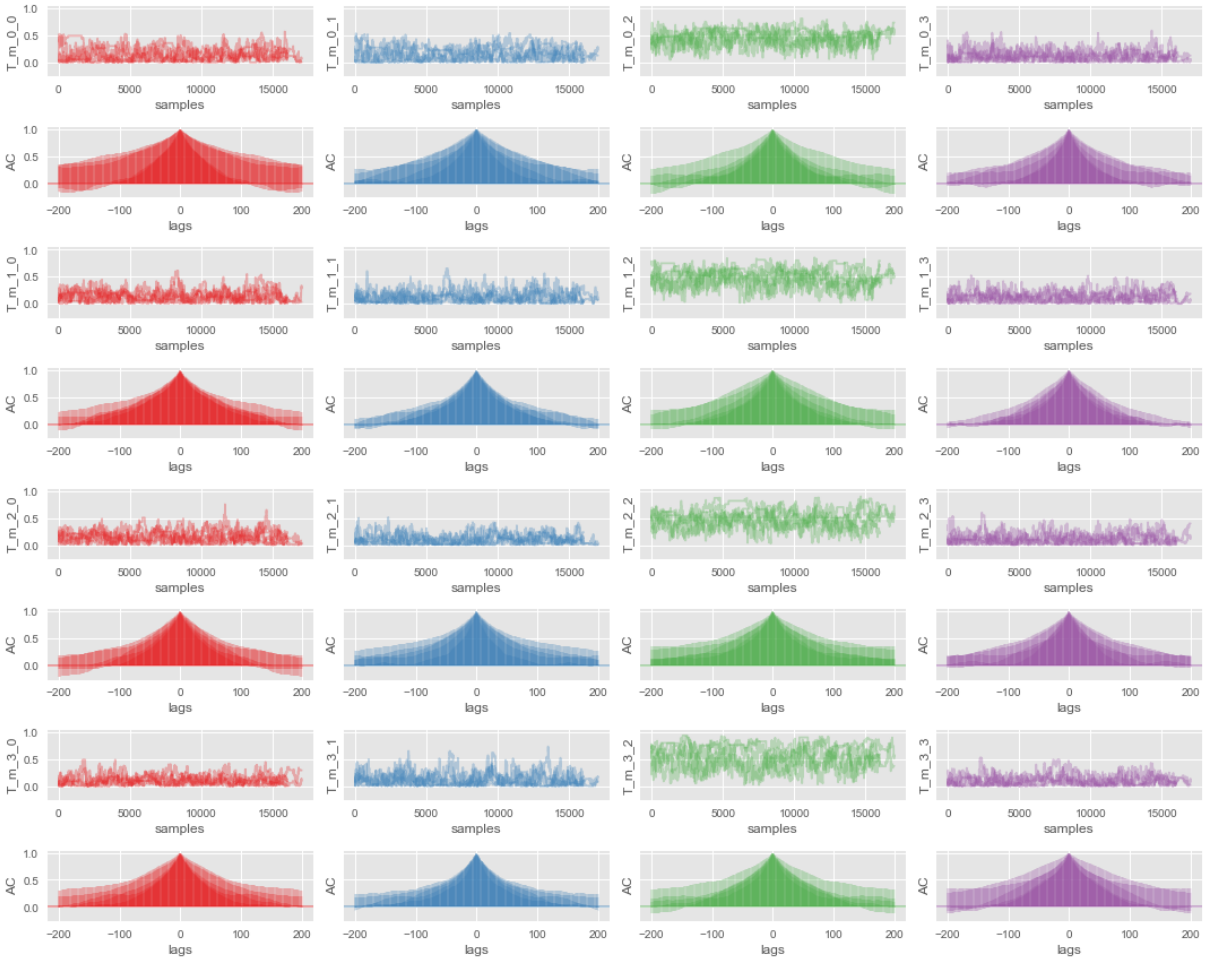


Figure D.5: MCMC chains and auto-correlation plots of T_z for individuals over sixty (Red: Eating and Drinking, Blue: Education and Health, Green: Retail, Purple: Sports and Entertainment. *2500 burn in applied*)

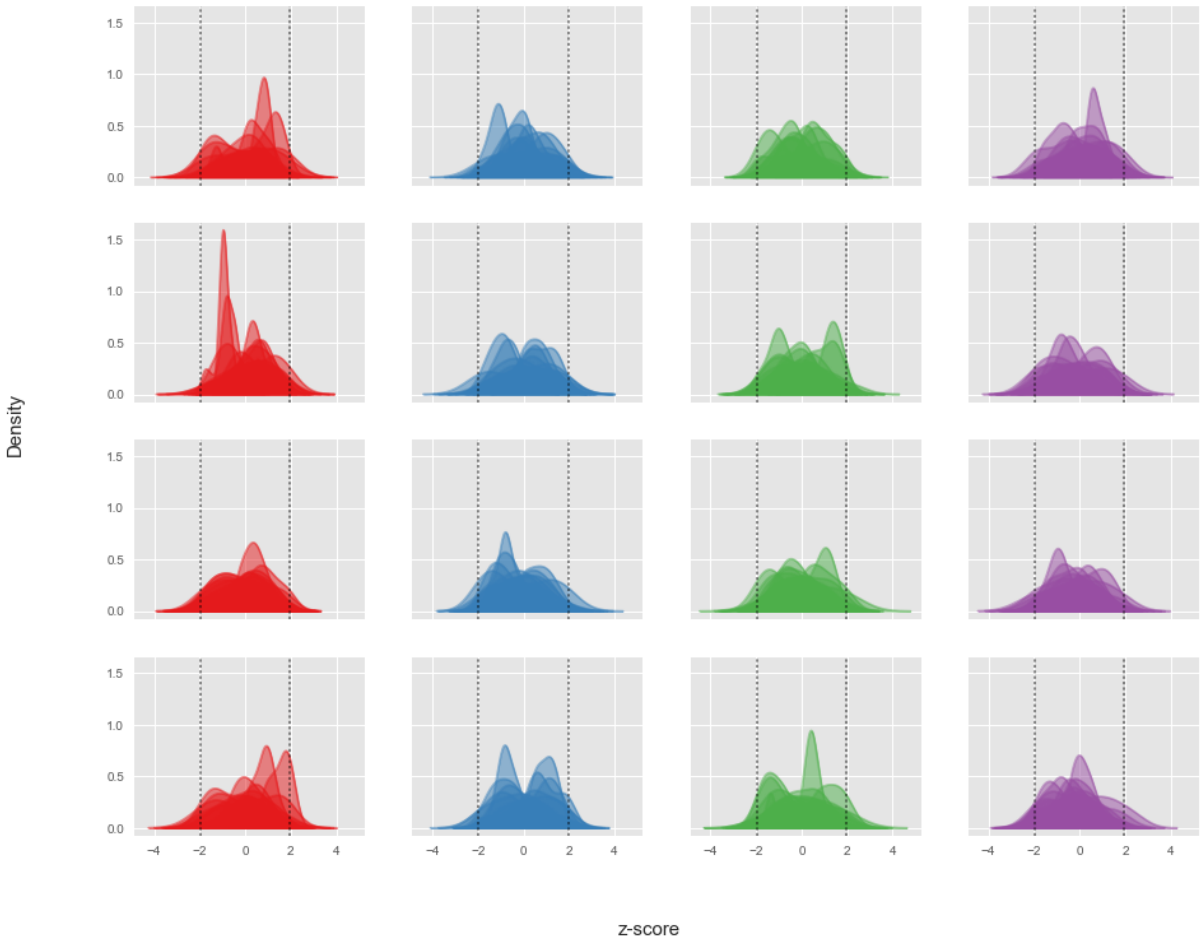


Figure D.6: Gewekes z-scores of T_z for all individuals over sixty years old.

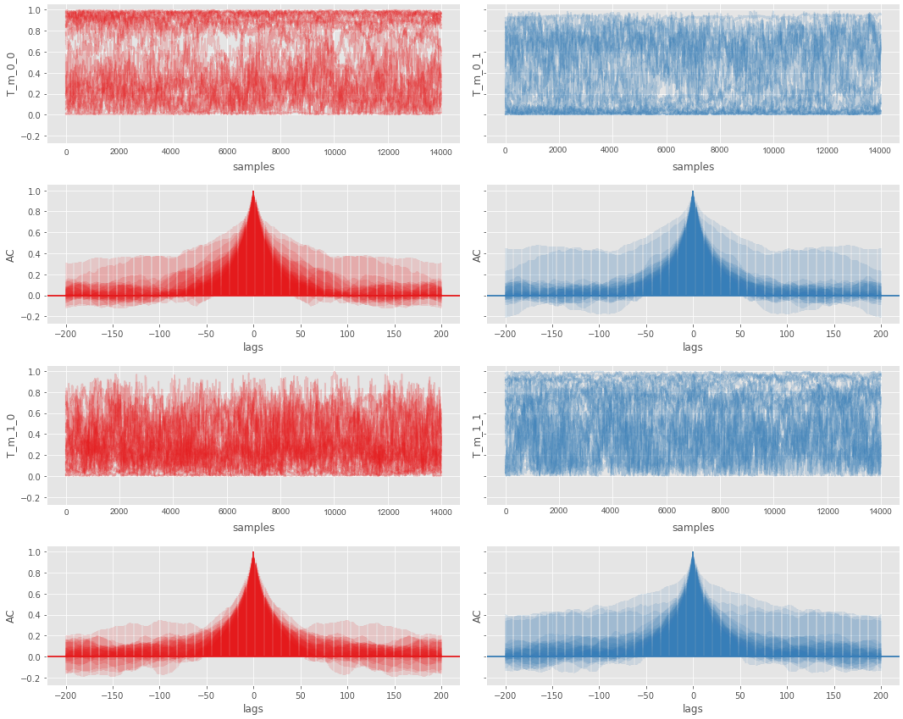


Figure D.7: MCMC chains and auto-correlation plots of T_m for individuals over sixty (Red: Bus, Blue: Rail. *6000 burn in applied*)

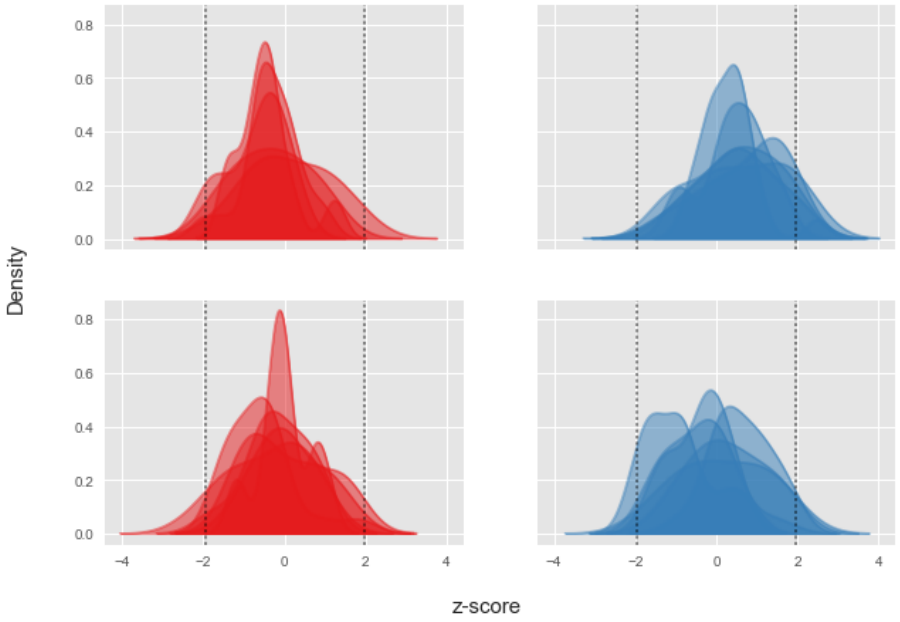


Figure D.8: Gewekes z-scores of T_m for all individuals over sixty years old

Next, the convergence diagnostics for the external factors ($\beta_{z,m}$), intercepts and internal covariates ($\alpha_{z,m}$) are presented in Figures D.9,D.10, D.11, D.12. Again, the effect of relatively high correlation of the external mobility covariates is evident from the slow mixing of the chain as illustrated by the autocorrelation plot in their regression coefficients.

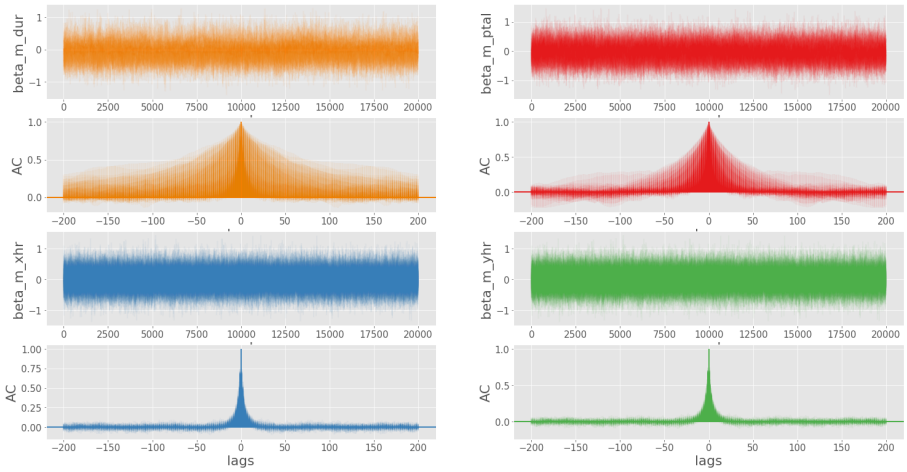


Figure D.9: MCMC chains and auto-correlation plots of β_m for individuals over sixty

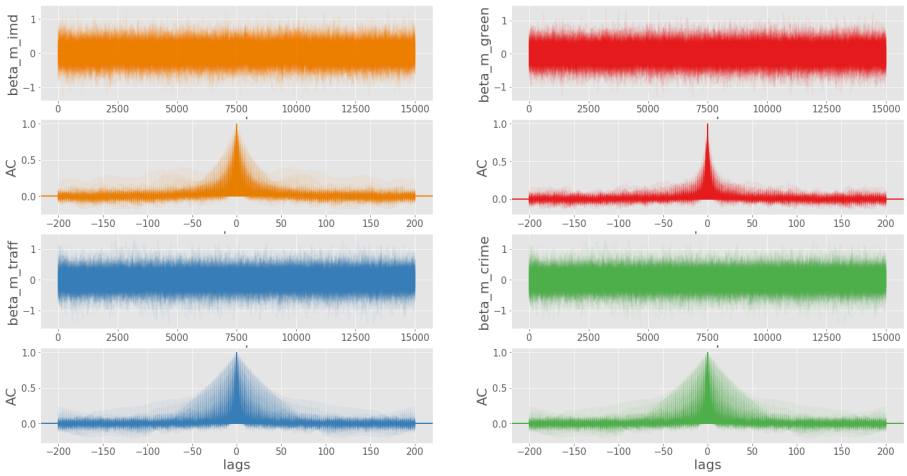


Figure D.10: MCMC chains and auto-correlation plots of β_z for individuals over sixty. *5000 burn in applied*

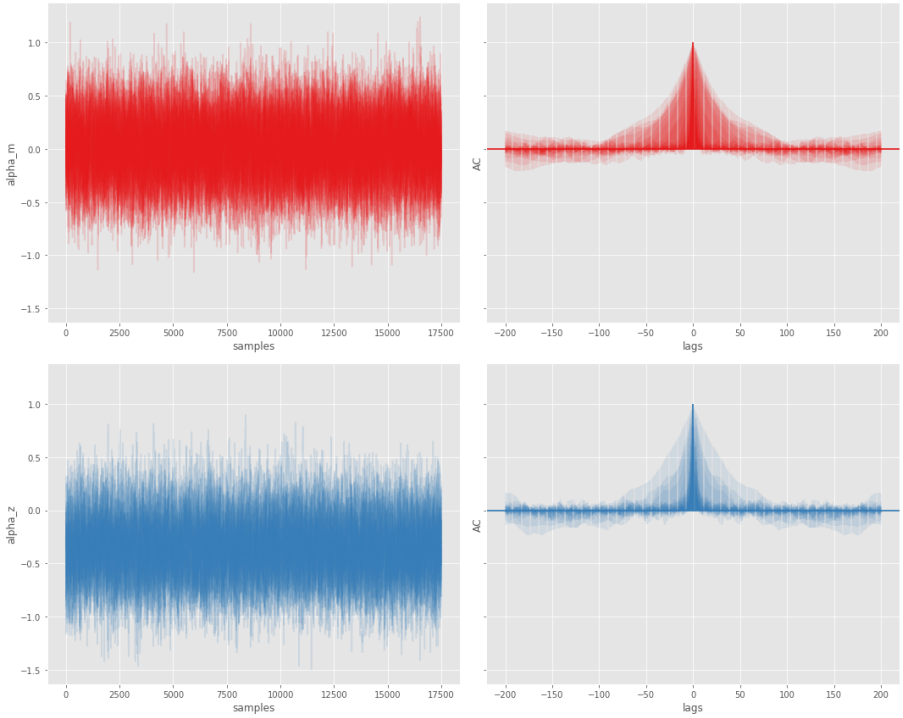


Figure D.11: MCMC chains and auto-correlation plots of $intercept_{m,z}$ for individuals over sixty (Red: mobility, Blue: activities). *2500 burn in applied*

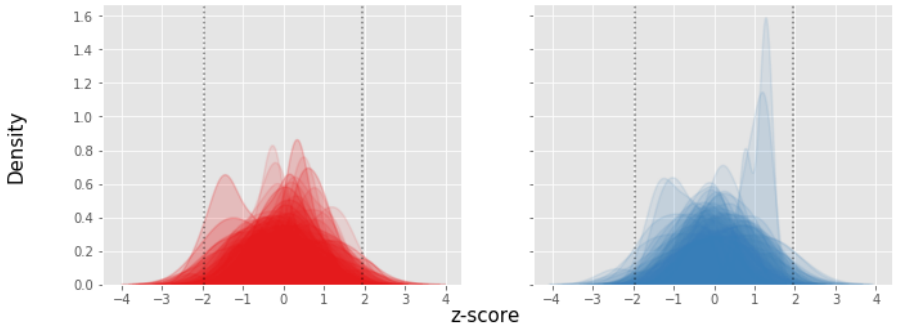


Figure D.12: Gewekes z-scores of $intercept_{m,z}$ for all individuals over sixty years old. (Red: mobility, Blue: activities)

The convergence results for the internal parameters modelled through the Dirichlet concentration parameters are presented in figures D.13, D.14, D.15 and D.16. Chain mixing for these nodes was generally acceptable judging from Geweke's z-scores and the trace and auto-correlation plots.

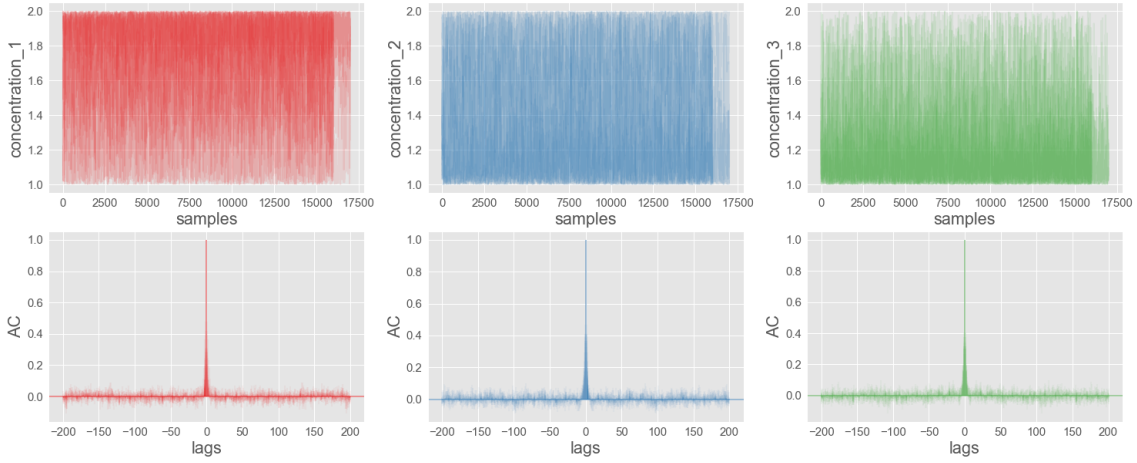


Figure D.13: MCMC chains and auto-correlation plots of α_m for individuals over sixty (Red: Bus, Blue: Rail, Green: Tram. *2500 burn in applied*)

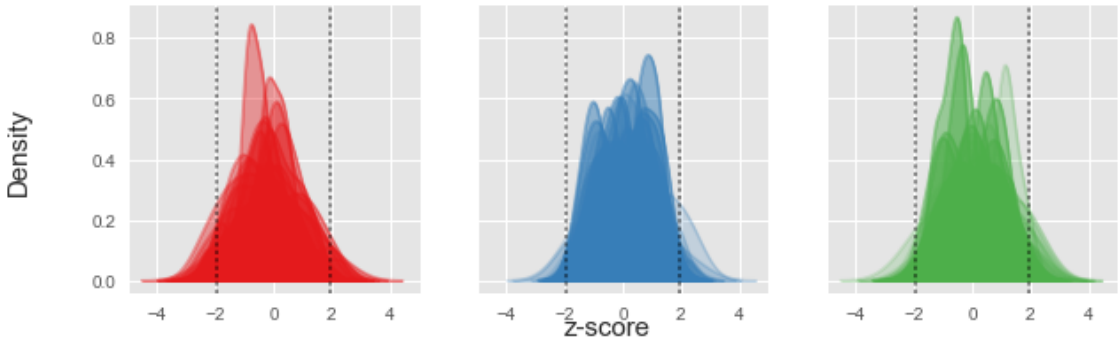


Figure D.14: Gewekes z-scores of α_m for all individuals over sixty years old. (Red: Bus, Blue: Rail, Green: Tram)

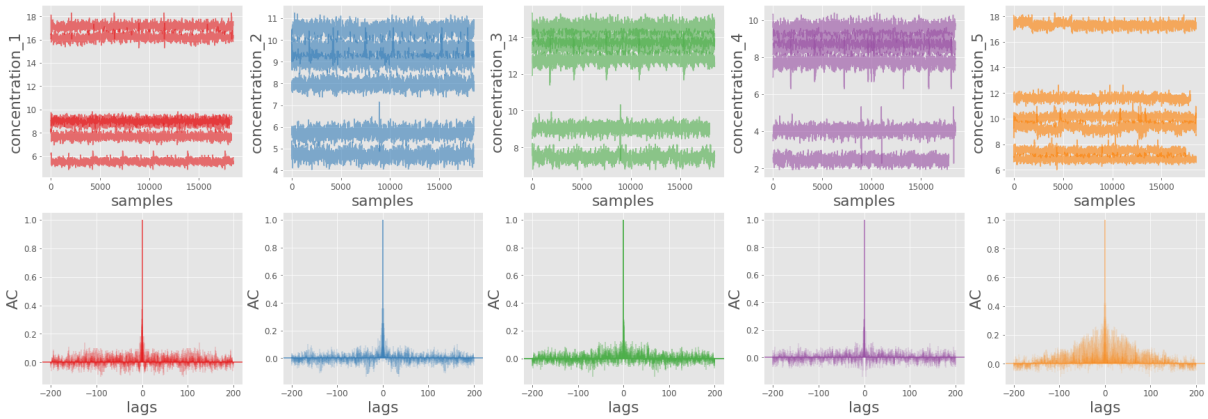


Figure D.15: MCMC chains and auto-correlation plots of α_z for individuals over sixty (Red: Eating and Drinking, Blue: Health and Education, Green: Retail, Purple: Sports and Entertainment, Yellow: Employment. *2500 burn in applied*)

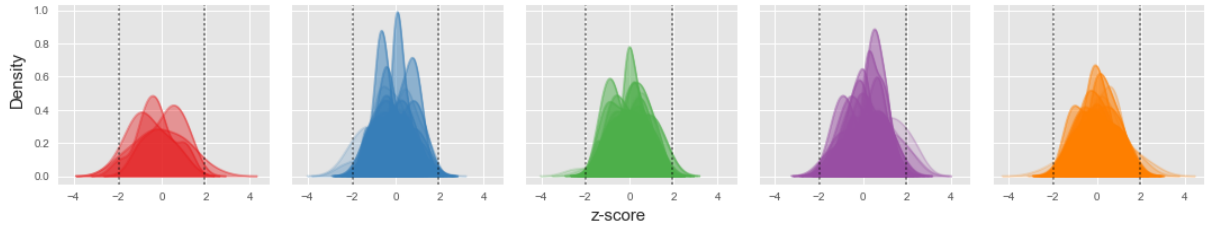


Figure D.16: Gewekes z-scores of α_z for all individuals over sixty years old. (Red: Eating and Drinking, Blue: Health and Education, Green: Retail, Purple: Sports and Entertainment, Yellow: Employment)

The convergence diagnostics for the rest of the population groups of interest, namely people with income lower than 15000£ and the unconstrained population sample is presented in appendix E.

D.1.2 Between chain convergence

Often, especially for over-complicated models, single long MCMC chain runs are considered inadequate to assess the convergence of the sampling procedure (Congdon 2007). Instead, multiple chains with random initial starting values are preferred to ensure that sampling is not "trapped" within a small region of the feature space. The Gelman-Rubin statistic (Gelman & Rubin 1992) is a diagnostic that uses an analysis of variance approach to check convergence of multiple chains. The general premise is that if convergence has been achieved, then the output of the chains will appear similar. The statistic uses the within and between chain sample variance, assessing their difference:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta_j} - \bar{\theta})$$
$$W = \frac{1}{m}\sum_{j=1}^{m}[\frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta_j})] \tag{D.1}$$

where B is the between variance, W the within chain variance and θ is each MCMC estimate.

These values are then used for an estimate of the marginal posterior variance:

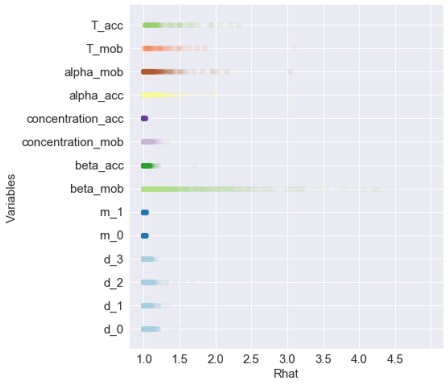
$$\hat{Var}(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}W \tag{D.2}$$

The Gelman-Rubin $\hat{R}$ statistic is then:

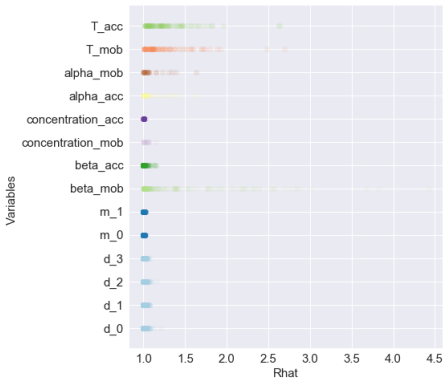
$$\hat{R} = \sqrt{\frac{\hat{Var}(\theta|y)}{W}} \tag{D.3}$$

In practice, $\hat{R}$ values close to one (with values around 1.5 as a rule of thumb) are desirable as an indication of convergence because as n goes to infinity the marginal posterior variance $\hat{Var}(\theta|y)$ and the within chain variance W will tend to be the same and coincide with the true variance of the estimate. The $\hat{R}$ value was computed for all stochastic nodes of the model for each individual using the 2×10000 sample chains. Figure D.17 below shows the results per population group:

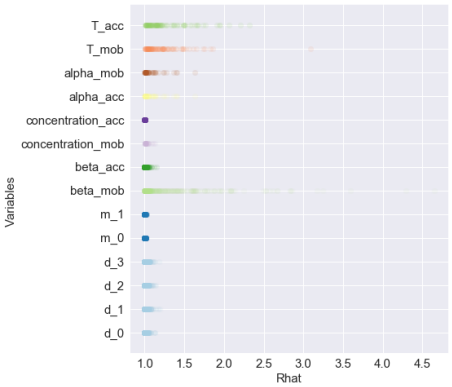
Convergence seems to have been achieved based on the $\hat{R}$ values (< 1.5) for the majority of variables for the three population groups of the Oyster card sample.



(a) $\hat{R}$ values for the unconstrained Oyster sample.



(b) $\hat{R}$ values for low income population group.



(c) $\hat{R}$ values for over sixty population group.

Figure D.17: Gelman-Rubin $\hat{R}$ values for the different population groups.

D.1.3 Activity type inference accuracy assessment

As already mentioned in section 6.3, a major limitation encountered in nearly all studies focusing on discovering urban activities using unlabeled trajectory data is the lack of ground truth information that could be used to verify experimental results. Instead, authors use secondary information such as travel surveys (Yin et al. 2018, Han & Sohn 2016, Alsger et al. 2018) to validate activity types related to non-commuting patterns or employ logic rules for activities such as home and

work (Wang et al. 2017). This is due to the unsupervised nature of inference, where learning of model parameters is performed without access to a reference dataset.

The limits of activity type inference accuracy using Bayesian networks and under different isochrone configuration settings has been explored in chapter 6.3. In particular, using Foursquare trajectory data, an overall accuracy of 56% was achieved for non-commuting to work activities, with activity detection results decreasing significantly for isochrone levels over 5min walking time. This provided an indication of the expected achievable accuracy levels using mobility datasets of low spatial resolution. However, it would be beneficial to perform a similar task with a similar dataset for this case study. For this task, a separate sample of 9 individuals was used to validate the activity type inference results of the CAA model (Sari Aslam et al. 2019). After accessing the Oyster card trajectory data from Transport for London, the volunteers were asked to label their data points with the activities performed at the vicinity of the transport access point. However, since the Oyster card records of the reference sample are not linked to iBus data, there was no possibility to determine the boarding and alighting bus stops. As a result, activity types are only relevant to rail services. Table D.1 below summarises the socioeconomic characteristics of the 9 volunteers.

ID	Sex	Age	Employment	Income	Ethnic group
79200886	M	30 to 40	Student	Below 25,000	British/Turkish
79200885	F	30 to 40	Full-time employment	Between 25000 to 40000	British
79200884	F	30 to 40	Full-time employment	Between 25000 to 40000	British/Colombian
79200881	F	20 to 30	Student	Below 25,000	Chinese
79200883	M	40 to 50	Full-time employment	Between 25000 to 40000	British-Chinese
79200887	F	30 to 40	Student	Below 25,000	Polish
79200882	M	20 to 30	Student	Below 25,000	Chinese
79200089	F	30 to 40	Full-time employment	Below 25,000	British/Turkish
79200888	F	30 to 40	Full-time employment	Below 25,000	British/Pakistani

Table D.1: Socioeconomic characteristics of volunteers

The information described above was then used to shape the prior distributions of activity types as described in section 6.4.1.1. For the purposes of this assessment, only the accessibility module of figure 6.35 was used. Inference was carried out by running two parallel MCMC chains of 5000 samples each. Figure D.18 below present the aggregated activity types confusion matrix for all volunteers in the sample. For the purposes of this assessment, the $argmax(P(d))$ was used as a benchmark for the inferred activity types. The overall accuracy for the 9 volunteers in the sample was 76%.

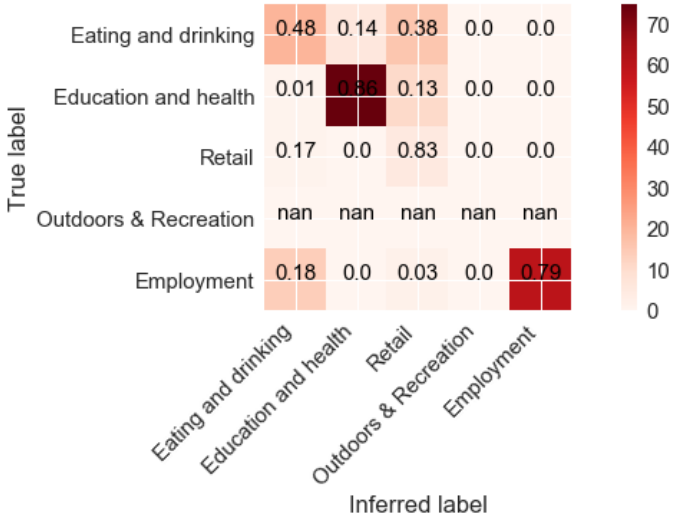


Figure D.18: Confusion matrix for inferred activities

As it can be seen from figure D.18, the results are consistent with the results presented in chapter 6.3, with the exception that in this case, the increased predictability of employment and education activities drive the overall accuracy higher. For the non commuting to work activities, the results are again characterised by a high overlap between *Eating and Drinking* and *Retail* category types.

Appendix E

Convergence diagnostics for people with income $< 15000£$, unconstrained population sample

E.1 people with income $< 15000£$

E.1.1 d node

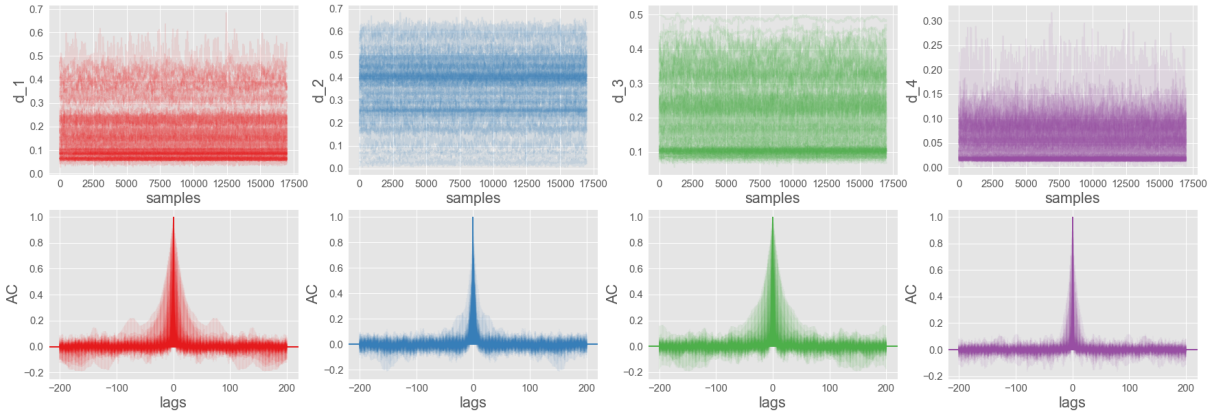


Figure E.1: MCMC traces and auto-correlation plots for the d nodes.

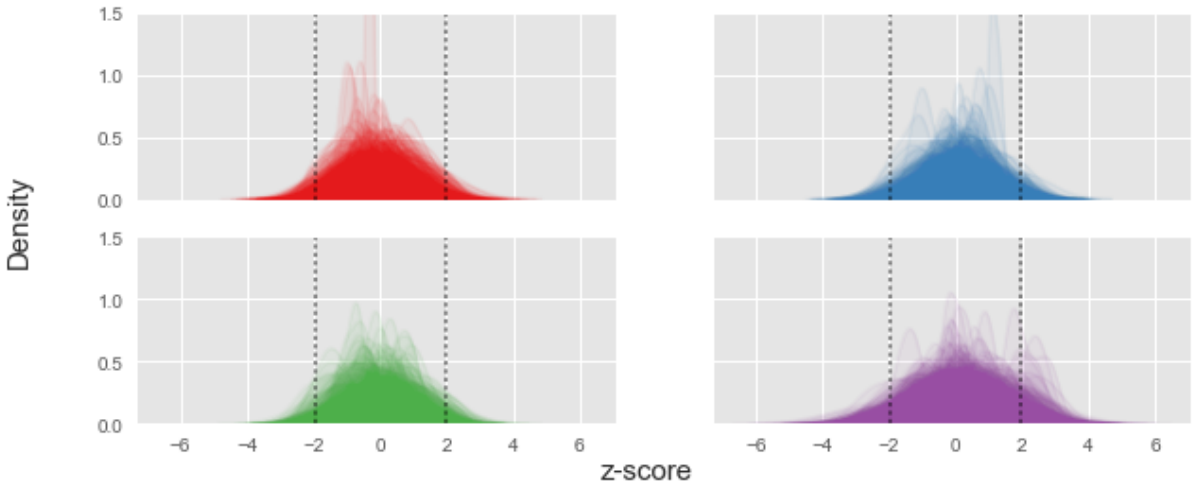


Figure E.2: Gewekes z-score plots for the d nodes.

E.1.2 m node

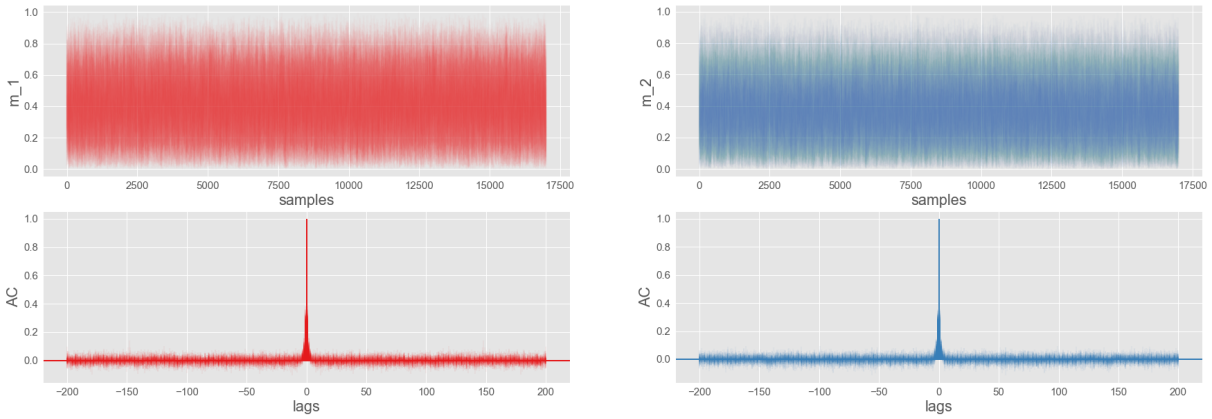


Figure E.3: MCMC traces and auto-correlation plots for the m nodes.

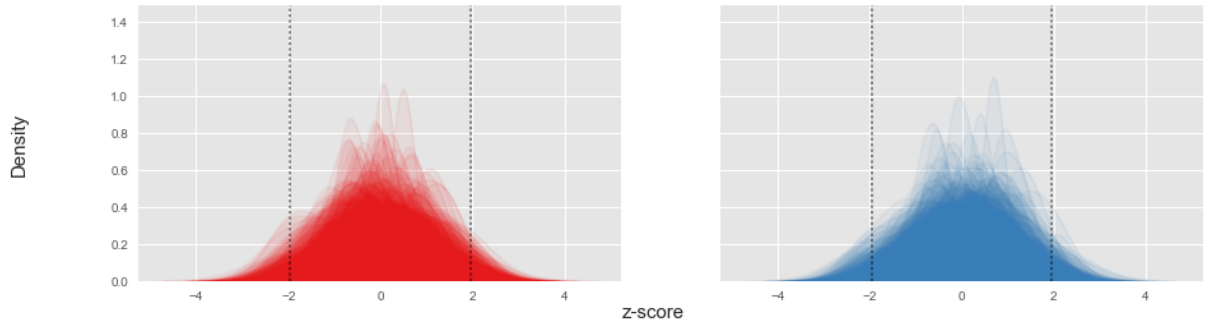


Figure E.4: Gewekes z-score plots for the m nodes.

E.1.3 T_z node

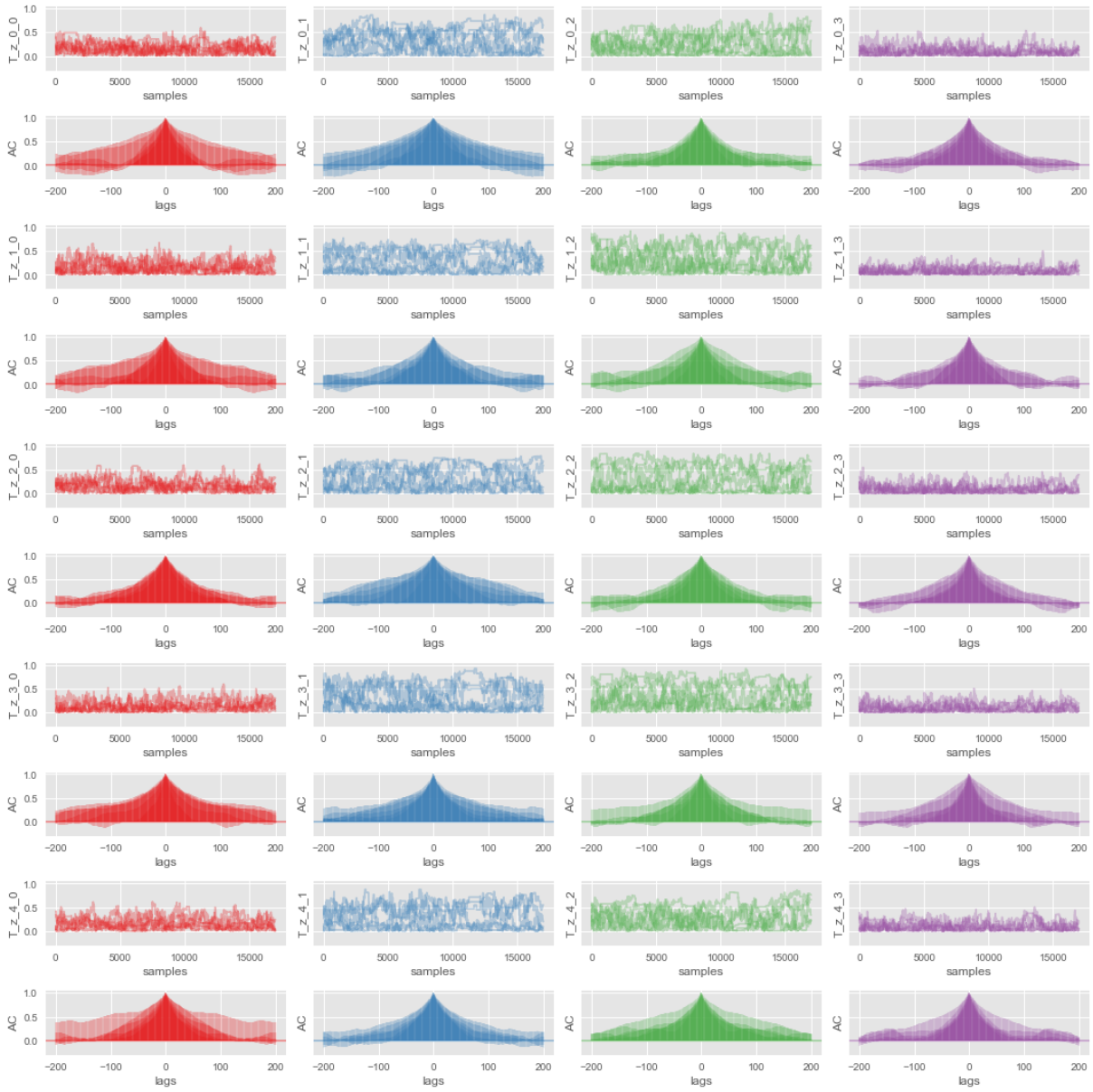


Figure E.5: MCMC traces and auto-correlation plots for the T_z node.

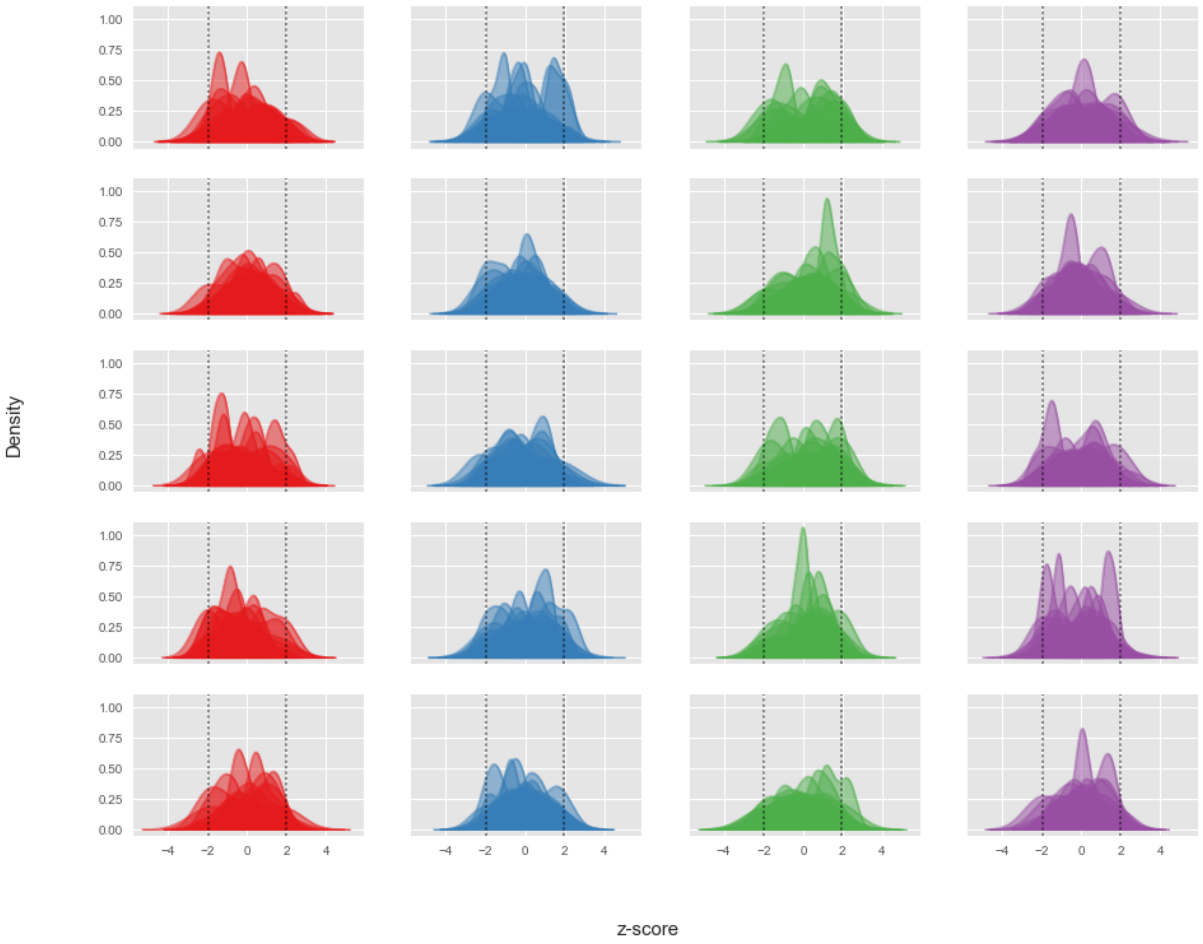


Figure E.6: Gewekes z-score plots for the T_z nodes.

E.1.4 T_m node

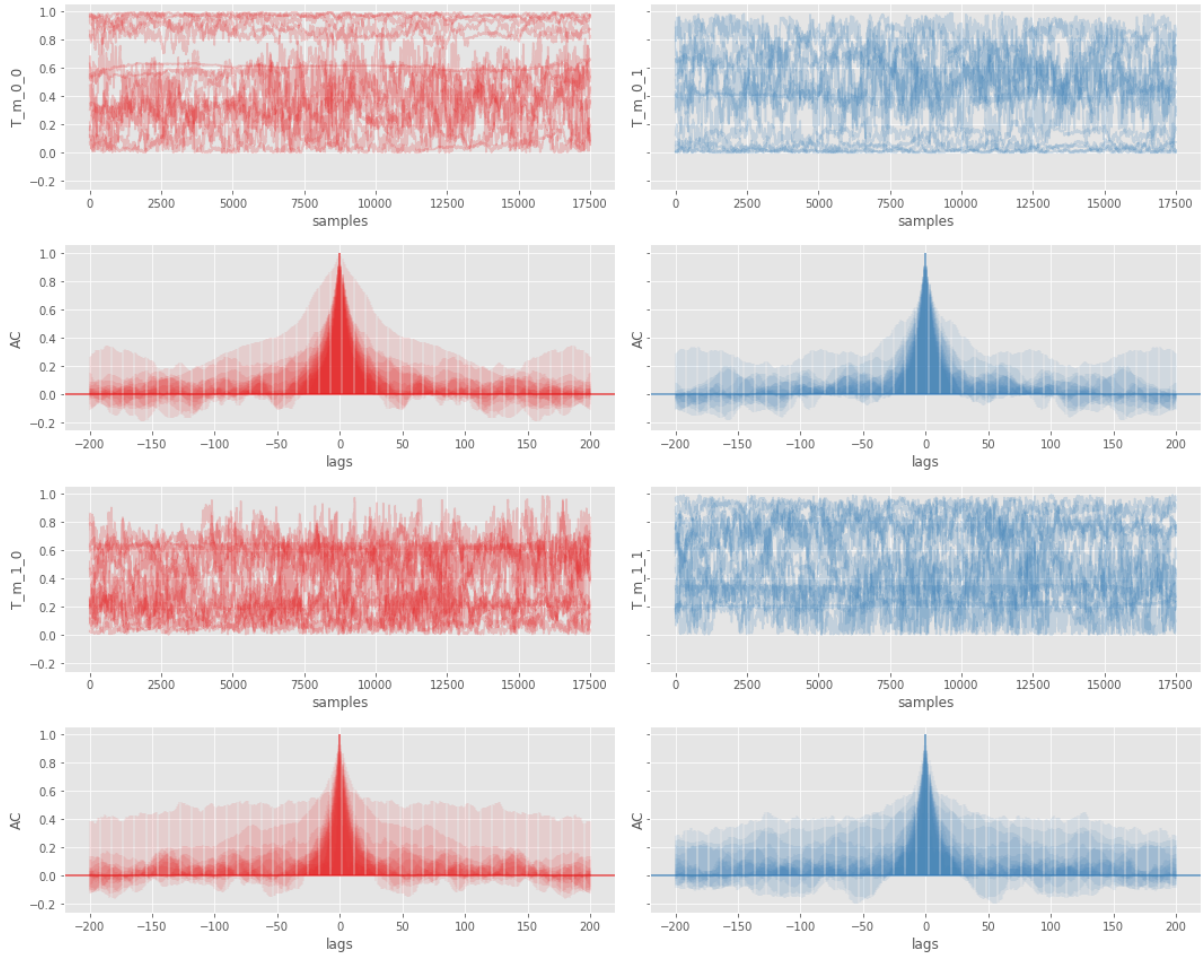


Figure E.7: MCMC traces and auto-correlation plots for the T_m node.

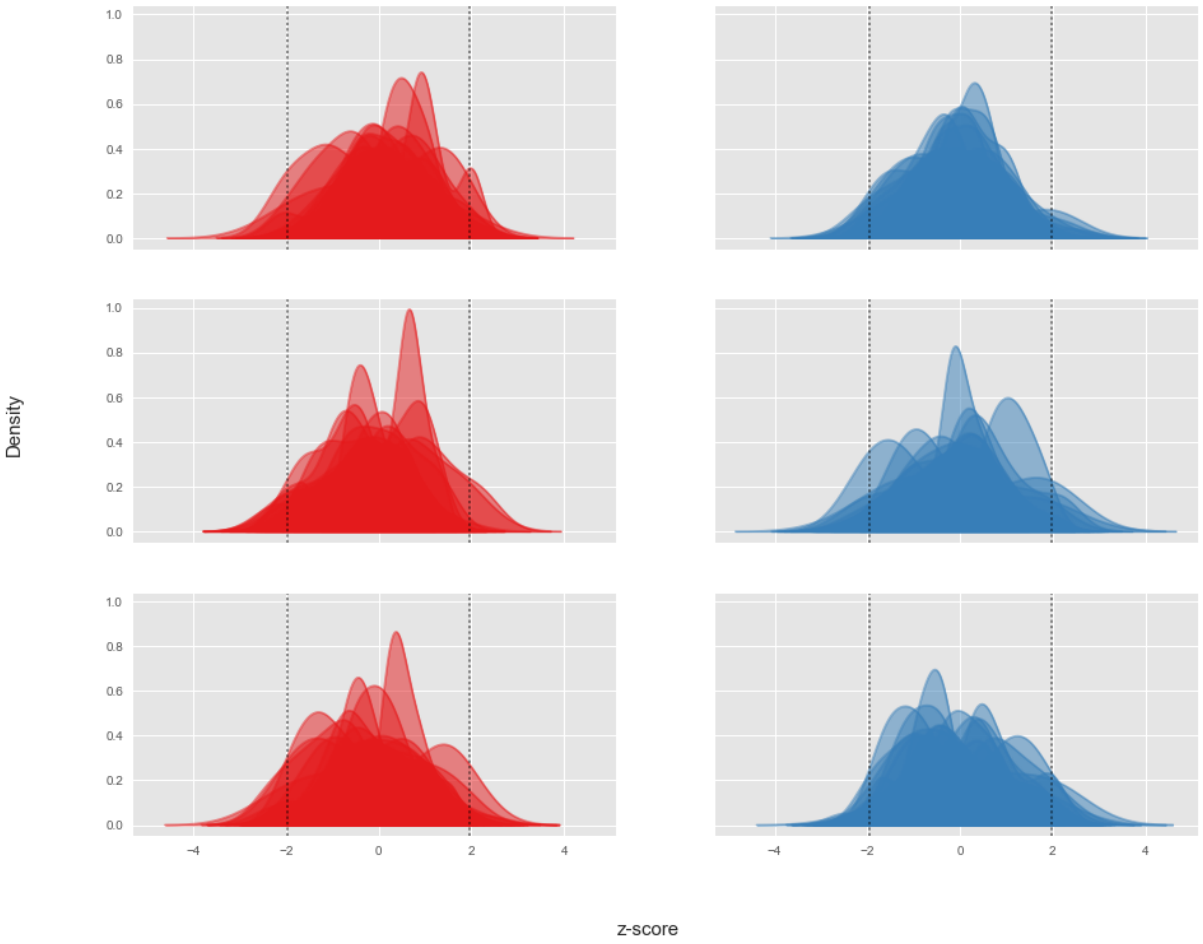


Figure E.8: Gewekes z-score plots for the T_m nodes.

E.1.5 β node

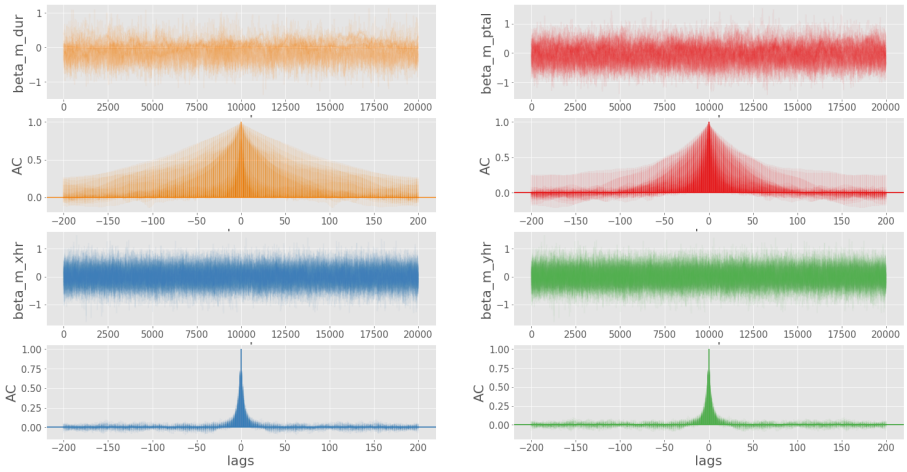


Figure E.9: MCMC traces and auto-correlation plots for the β_m node.

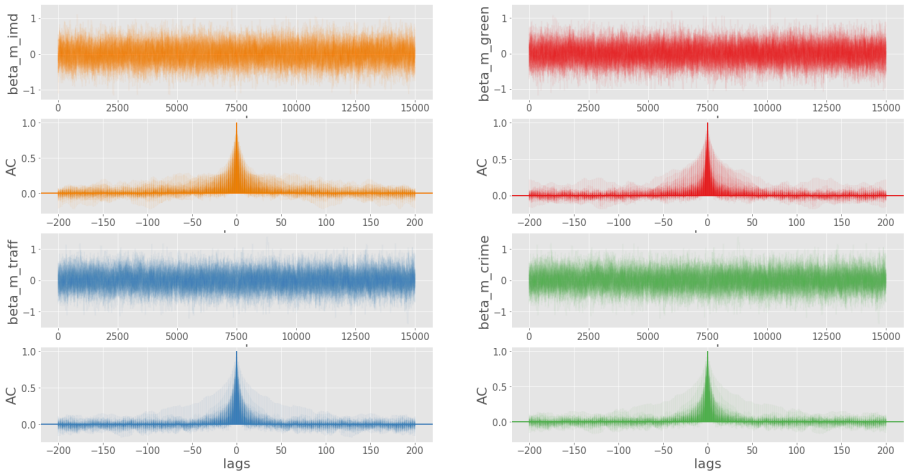


Figure E.10: MCMC traces and auto-correlation plots for the β_{acc} node.

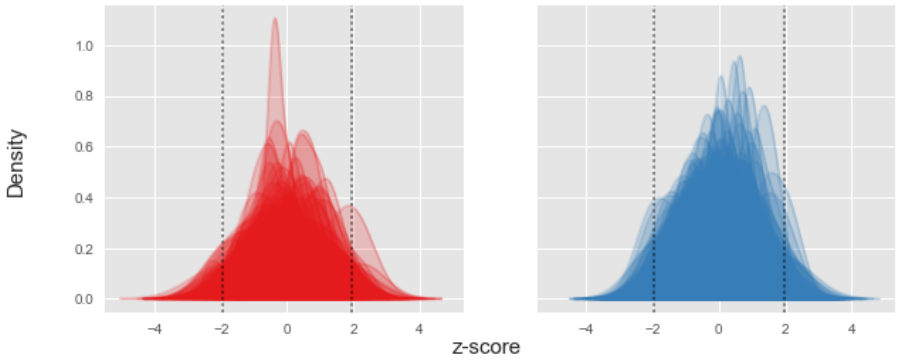


Figure E.11: Gewekes z-score plots for the β nodes.

E.1.6 *intercept* node

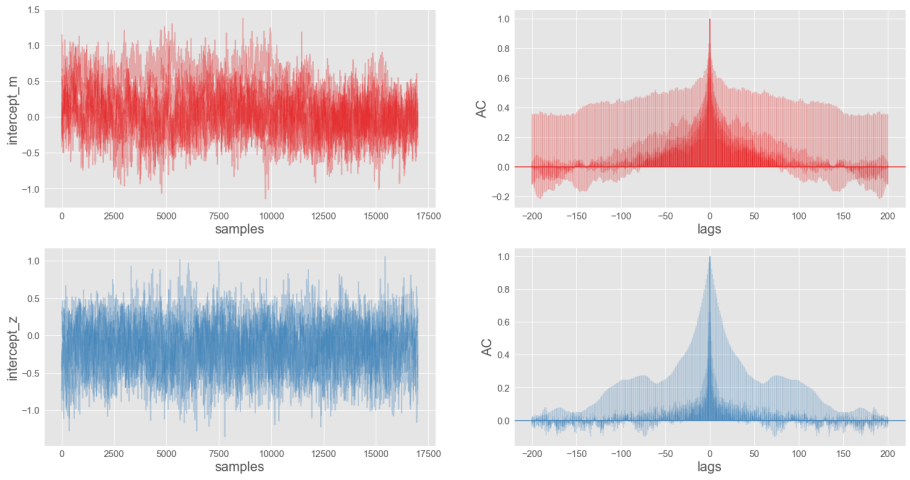


Figure E.12: MCMC traces and auto-correlation plots for the *intercept* nodes.

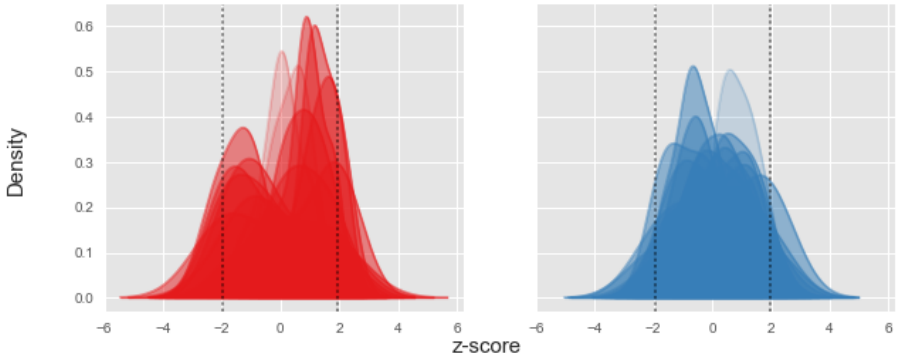


Figure E.13: Gewekes z-score plots for the *intercept* nodes.

E.1.7 α node

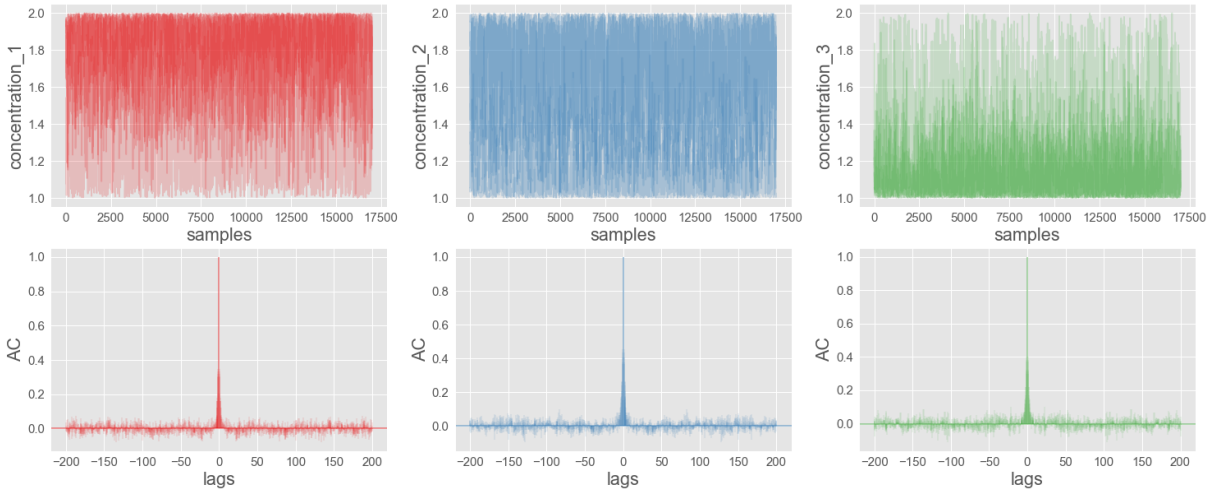


Figure E.14: MCMC traces and auto-correlation plots for the α_m nodes.

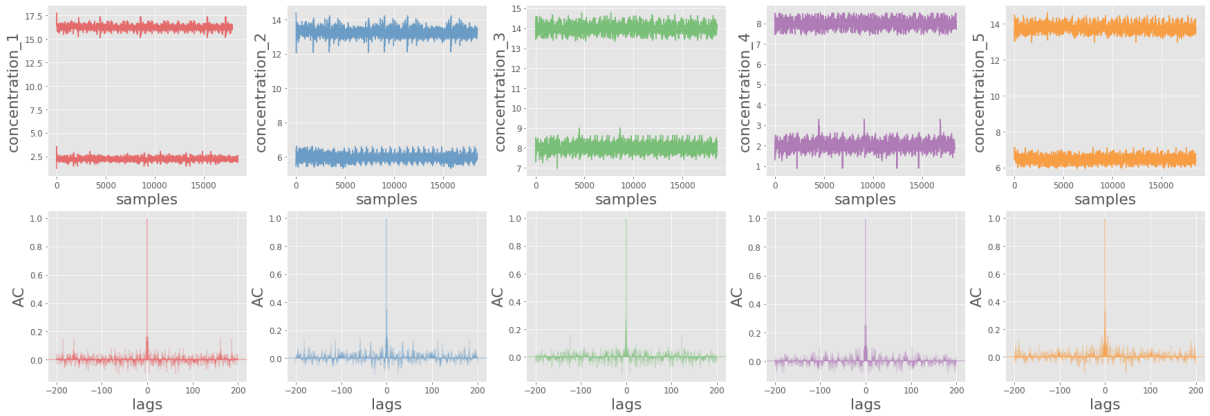


Figure E.15: MCMC traces and auto-correlation plots for the α_z nodes.

E.2 Unconstrained sample

E.2.1 d node

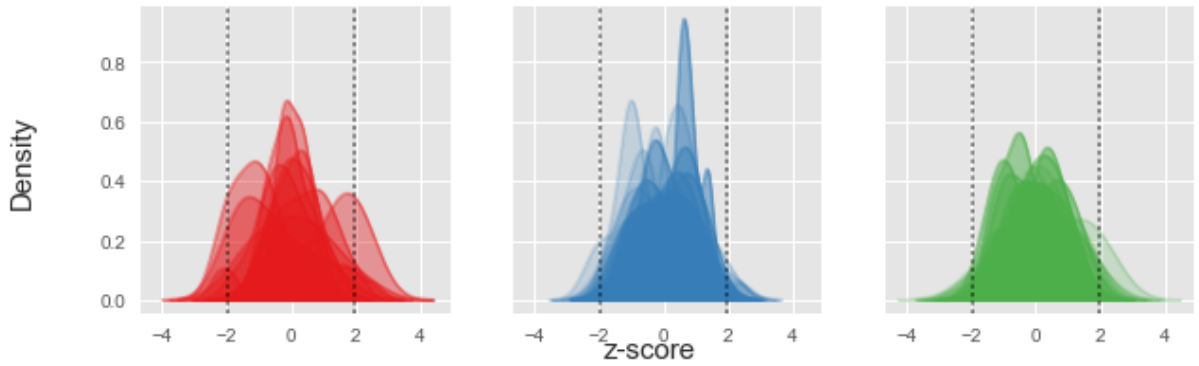


Figure E.16: Gewekes z-score plots for the α_m nodes.

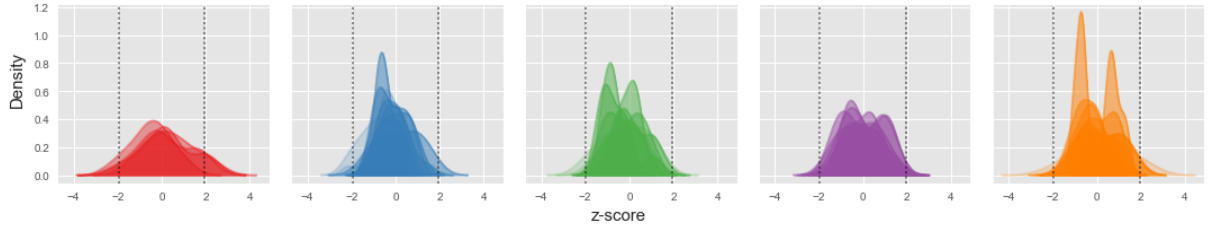


Figure E.17: Gewekes z-score plots for the α_z nodes.

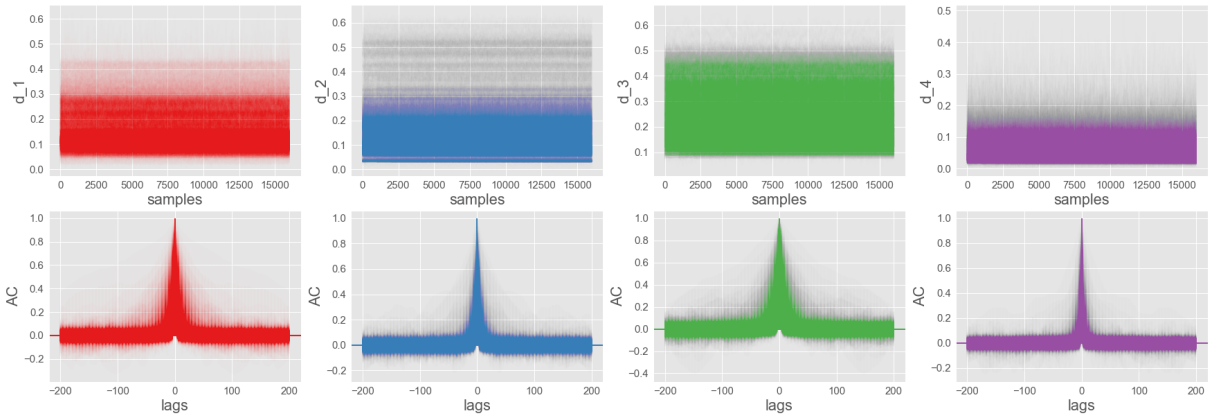


Figure E.18: MCMC traces and auto-correlation plots for the d nodes.

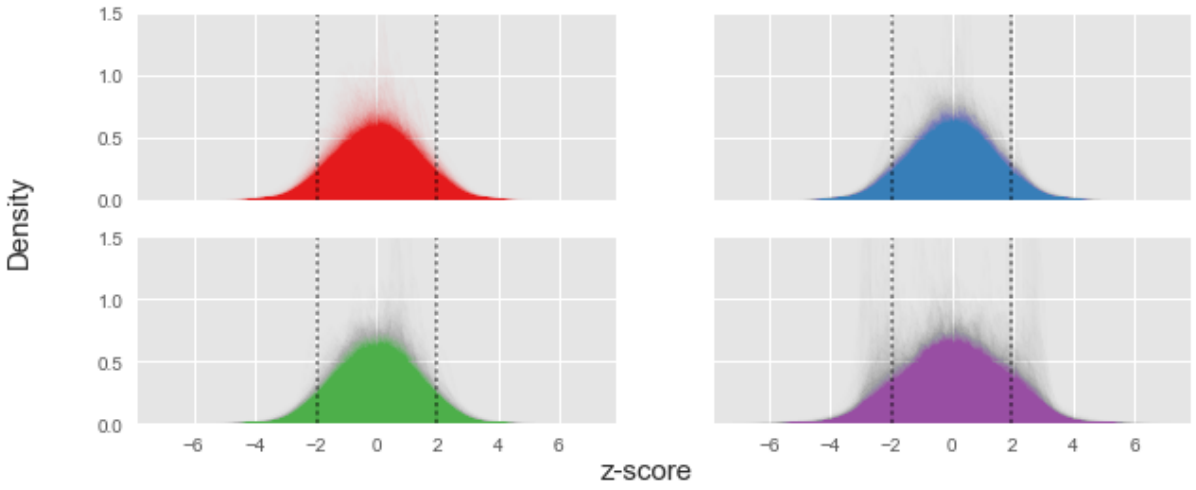


Figure E.19: Gewekes z-score plots for the d nodes.

E.2.2 m node

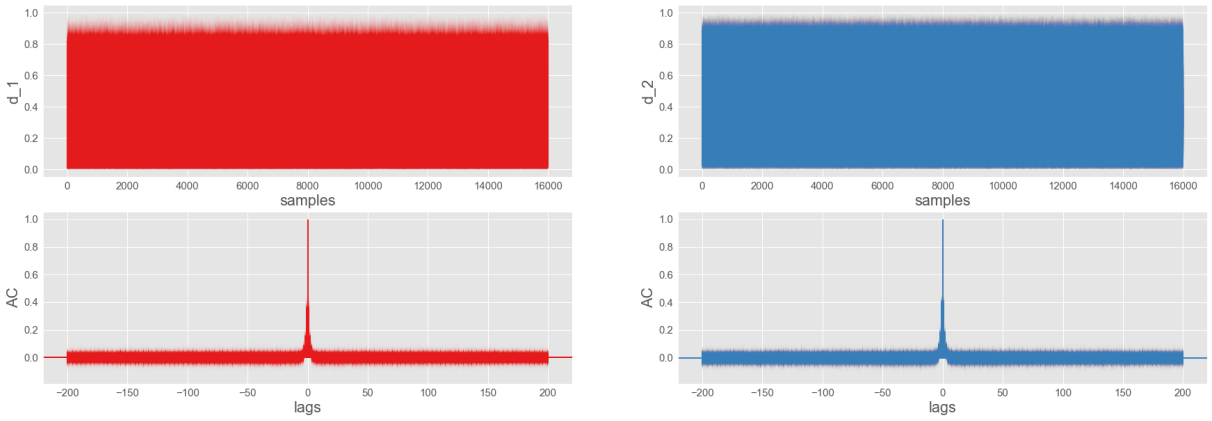


Figure E.20: MCMC traces and auto-correlation plots for the m nodes.

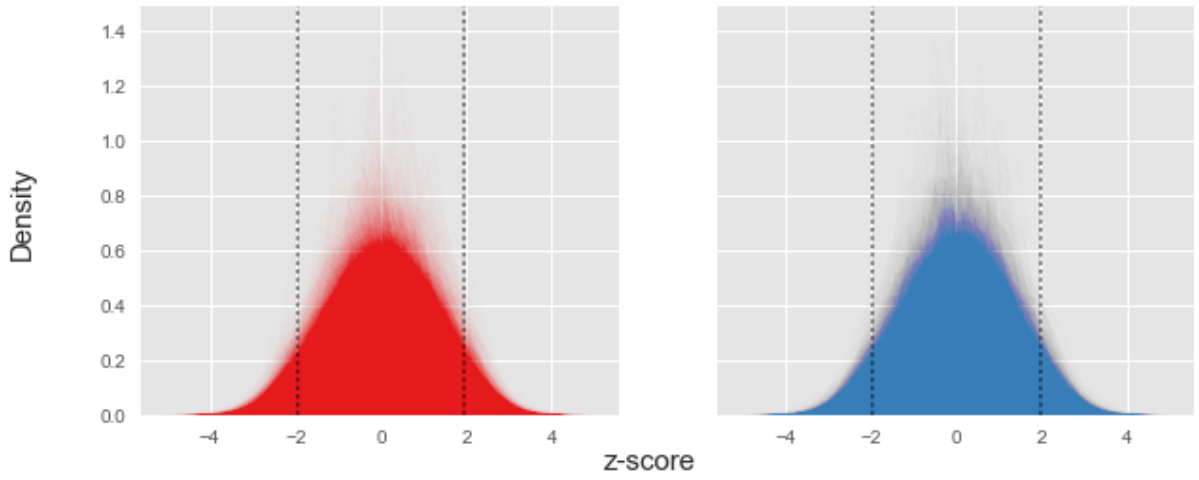


Figure E.21: Gewekes z-score plots for the m nodes.

E.2.3 T_z node

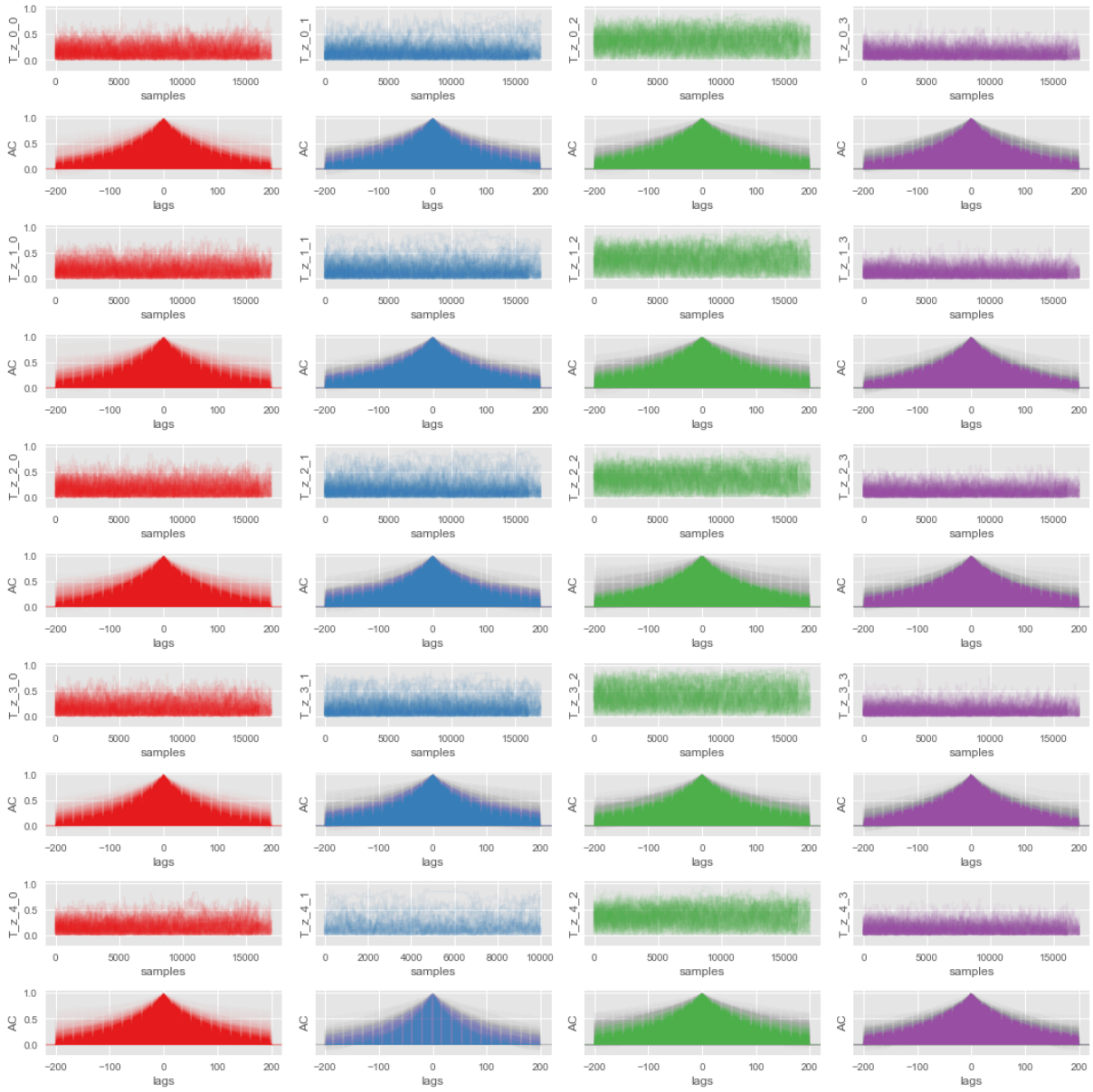


Figure E.22: MCMC traces and auto-correlation plots for the T_z node.

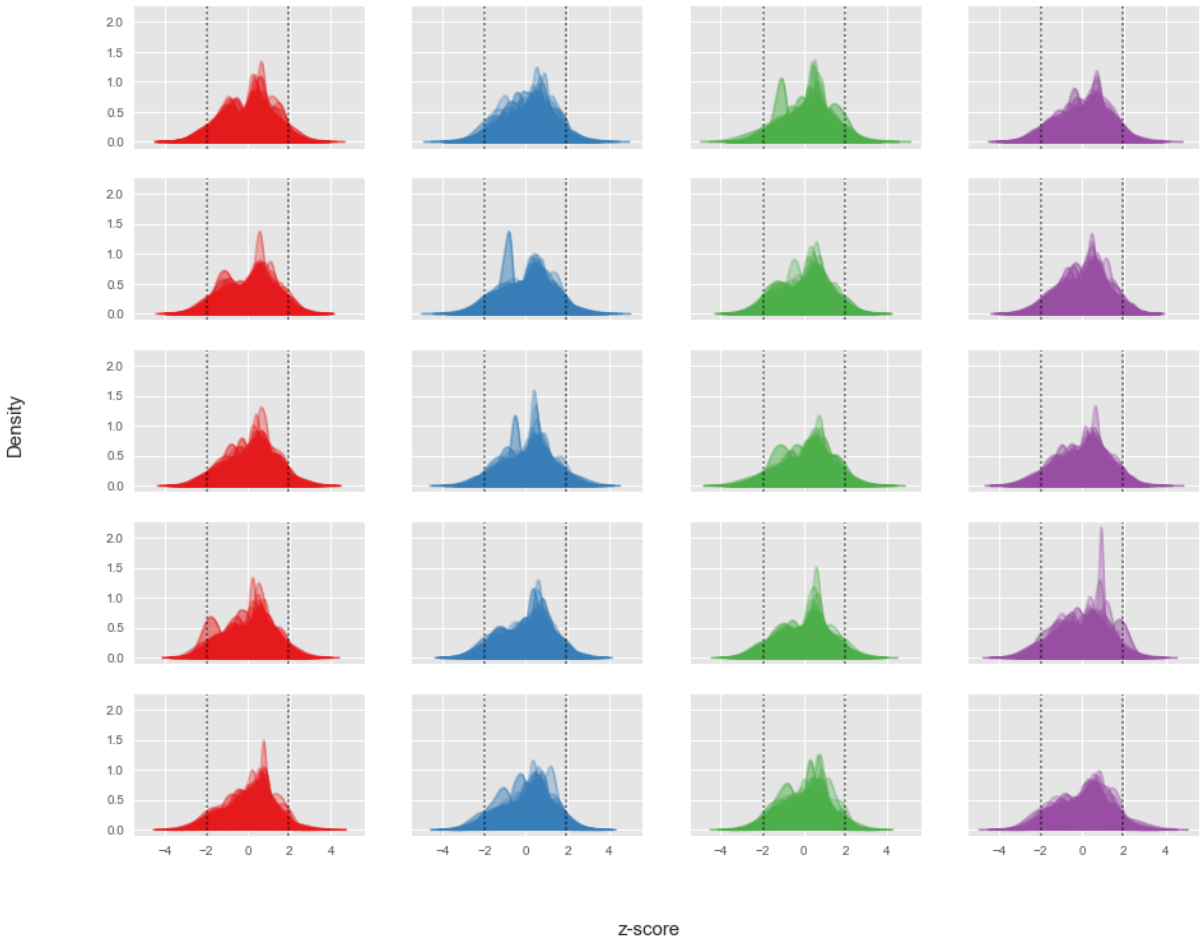


Figure E.23: Gewekes z-score plots for the T_z nodes.

E.2.4 T_m node

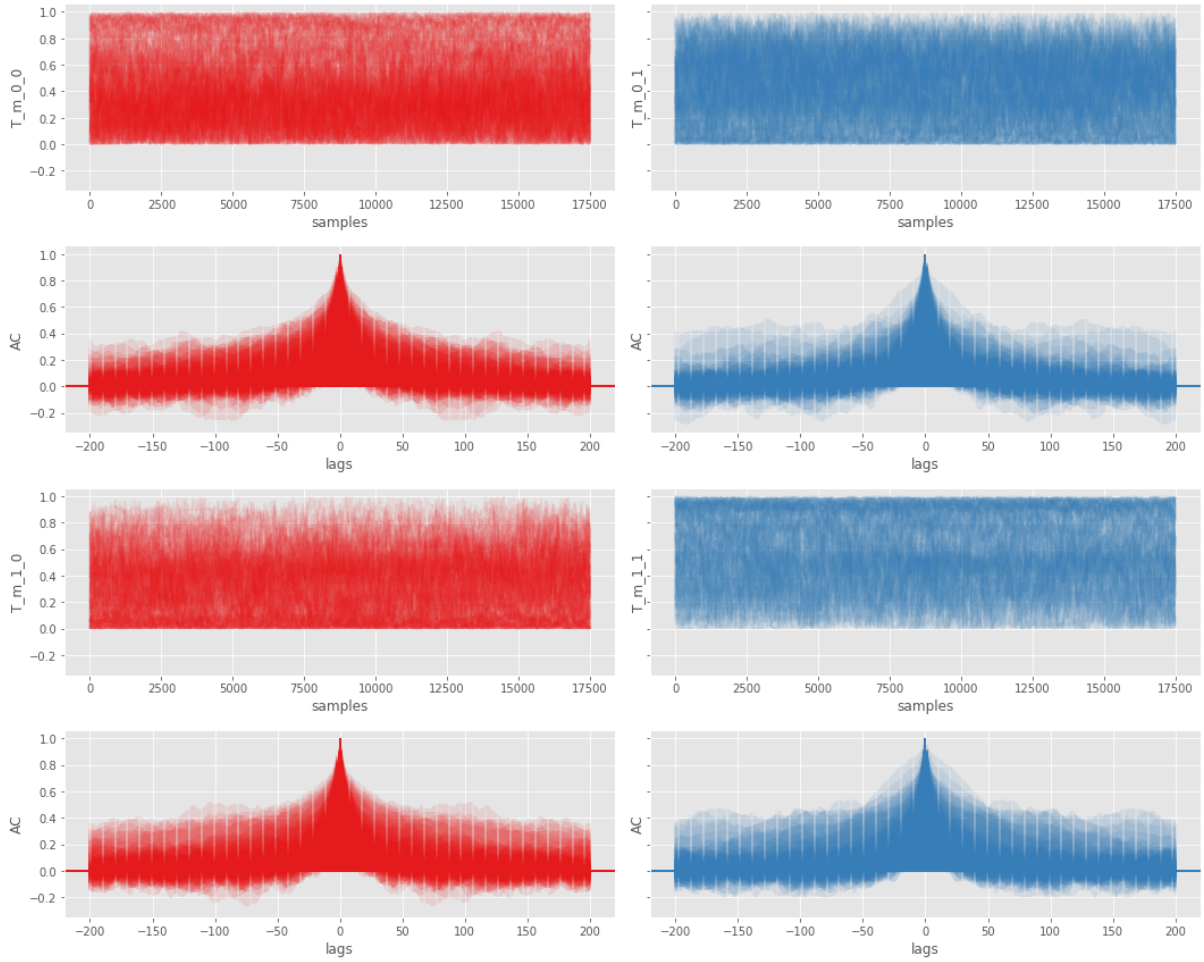


Figure E.24: MCMC traces and auto-correlation plots for the T_m node.

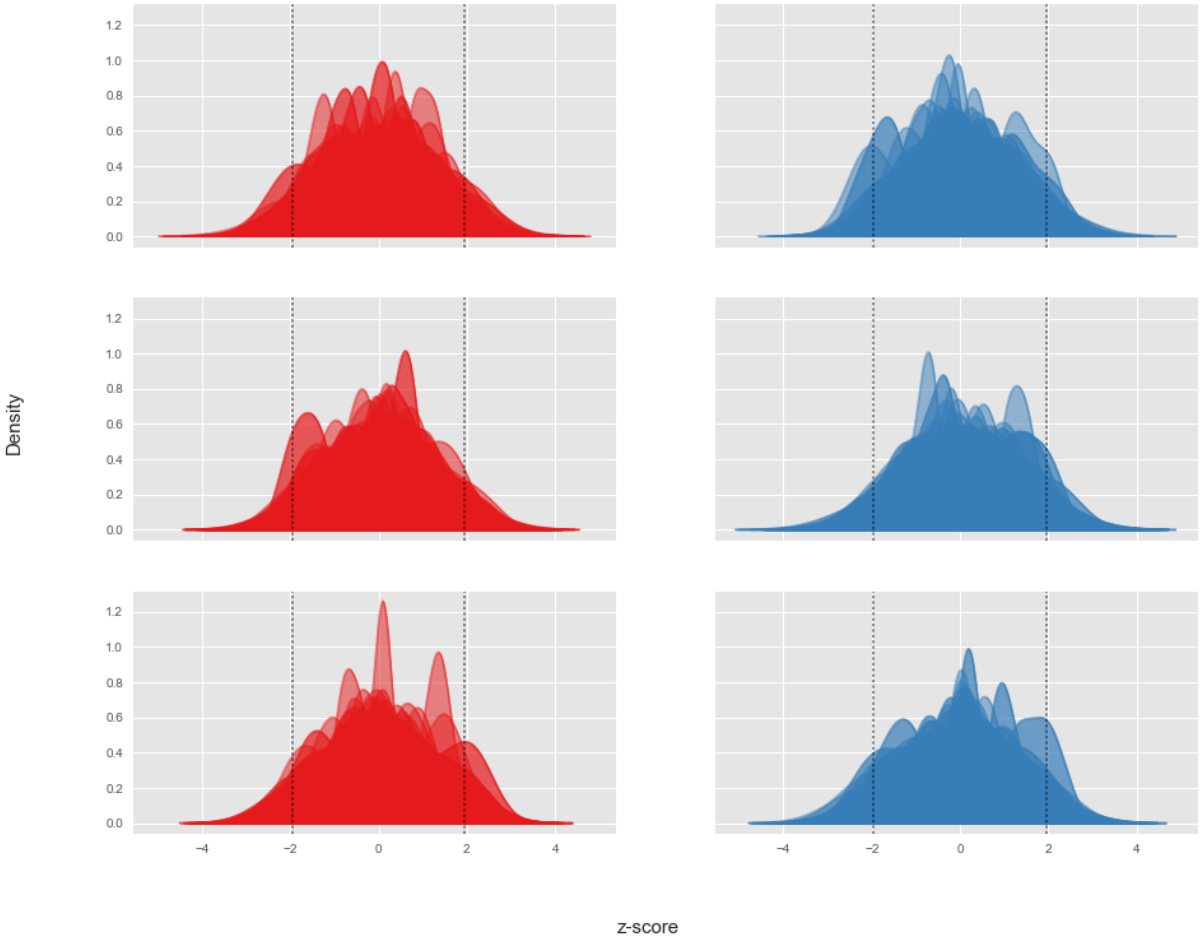


Figure E.25: Gewekes z-score plots for the T_m nodes.

E.2.5 β node

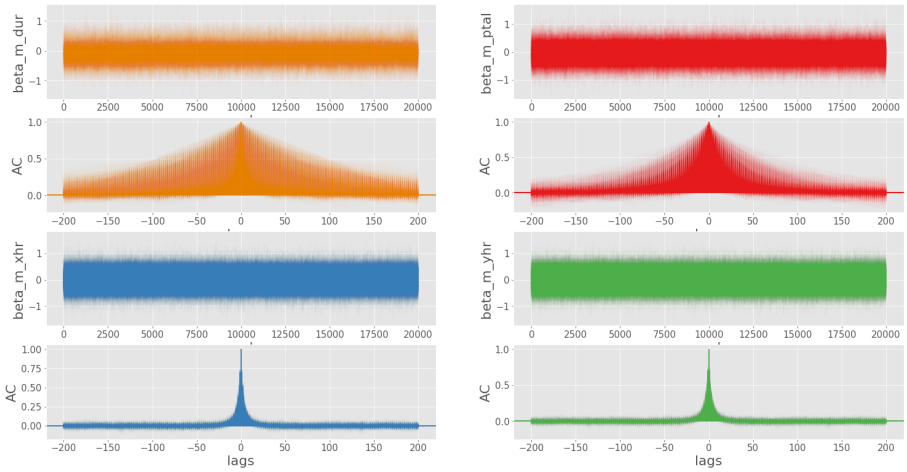


Figure E.26: MCMC traces and auto-correlation plots for the β_m node.

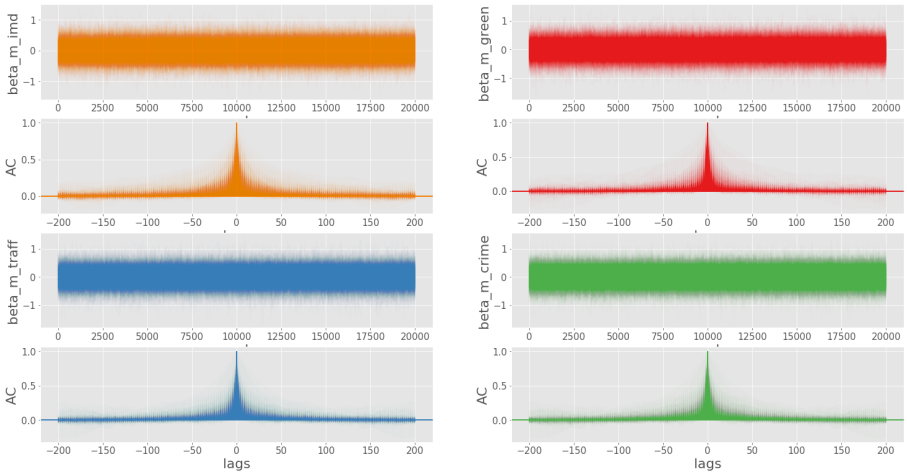


Figure E.27: MCMC traces and auto-correlation plots for the β_{acc} node.

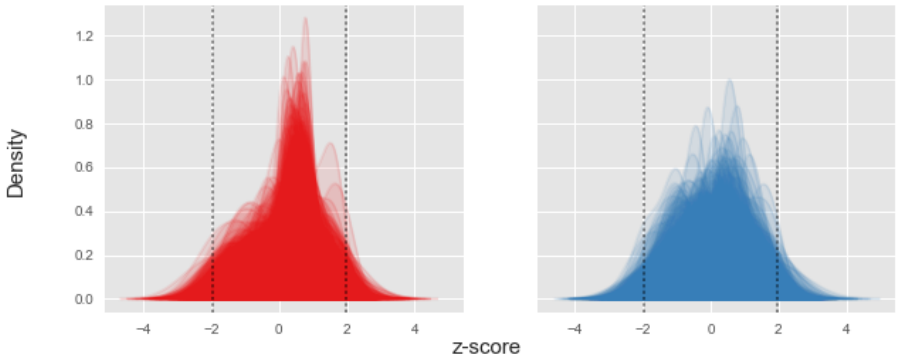


Figure E.28: Gewekes z-score plots for the β nodes.

E.2.6 *intercept* node

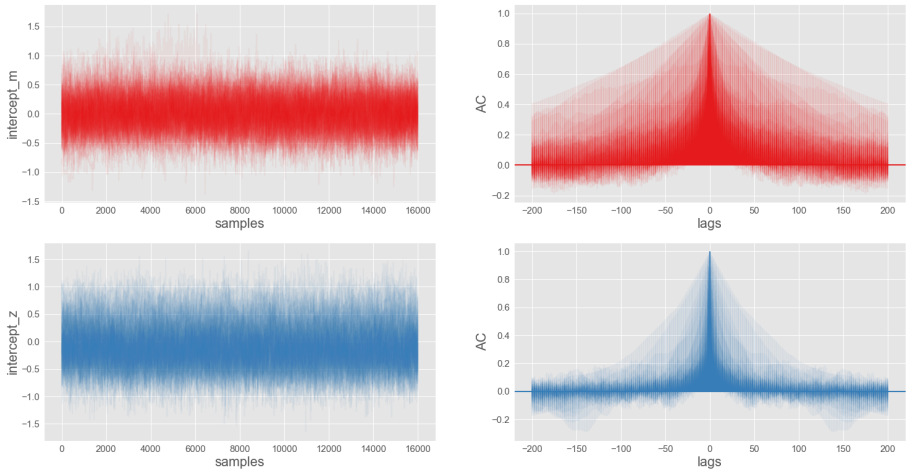


Figure E.29: MCMC traces and auto-correlation plots for the *intercept* nodes.

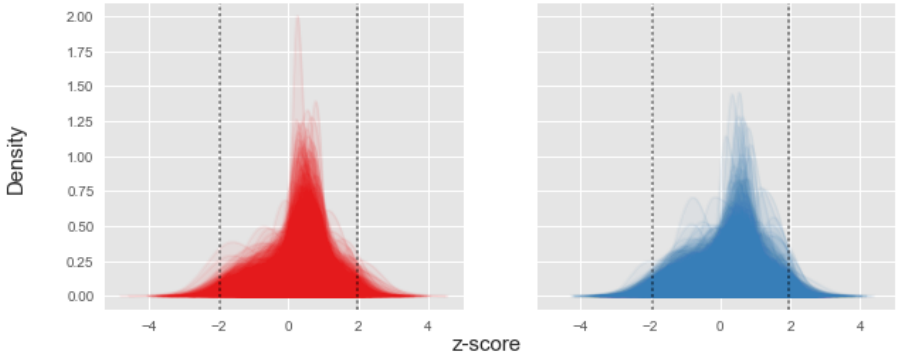


Figure E.30: Gewekes z-score plots for the *intercept* nodes.

E.2.7 α node

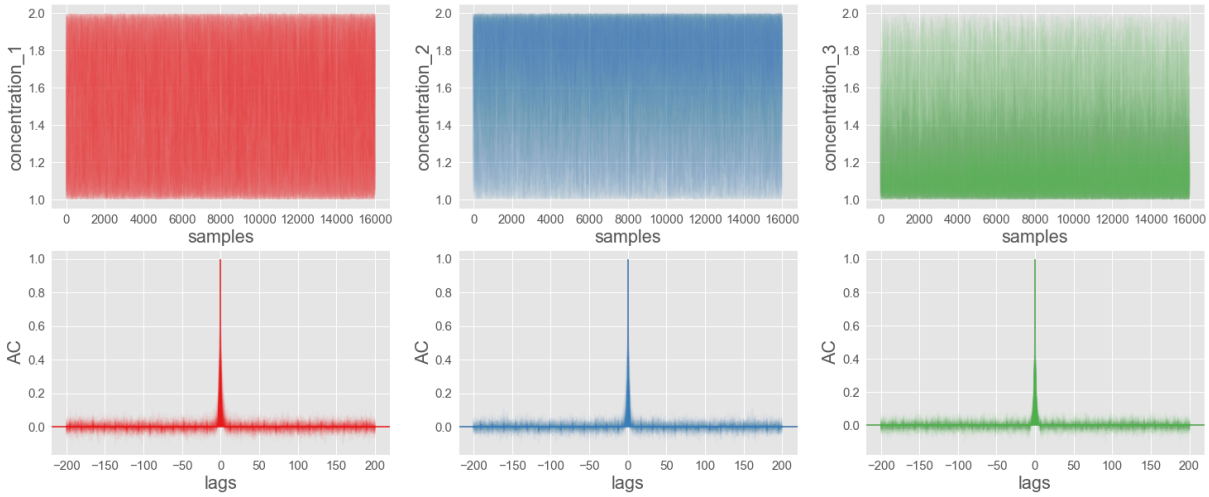


Figure E.31: MCMC traces and auto-correlation plots for the α_m nodes.

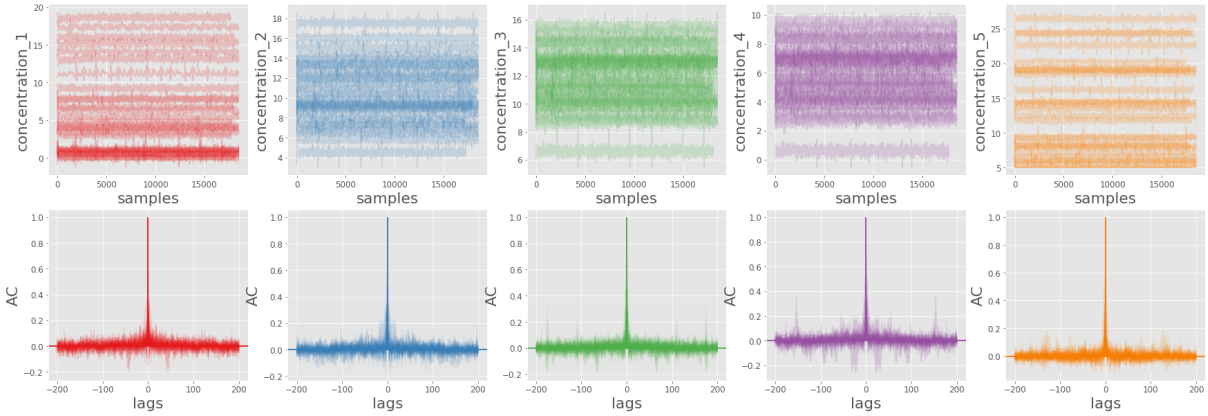


Figure E.32: MCMC traces and auto-correlation plots for the α_z nodes.

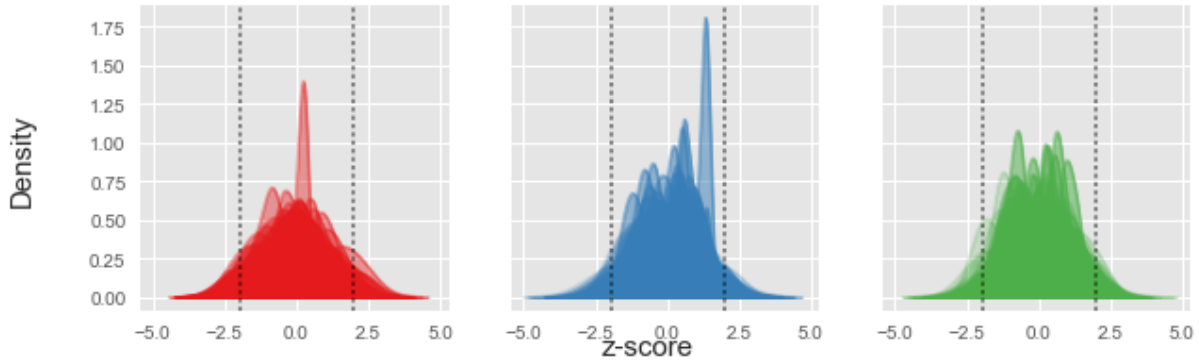


Figure E.33: Gewekes z-score plots for the α_m nodes.

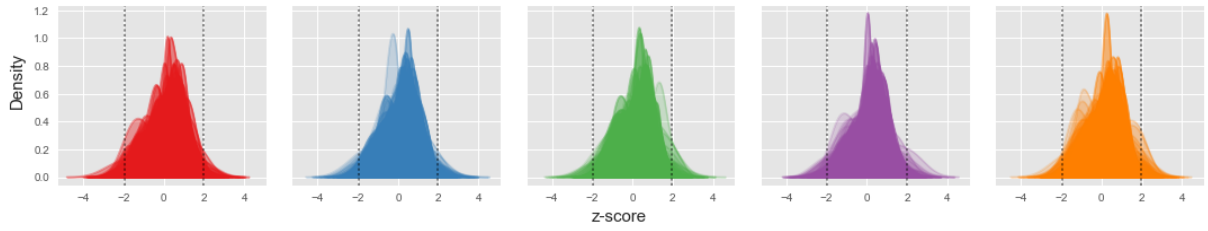


Figure E.34: Gewekes z-score plots for the α_z nodes.

Appendix F

Ethics forms

IMPORTANT: ALL FIELDS MUST BE COMPLETED. THE FORM SHOULD BE COMPLETED IN PLAIN ENGLISH UNDERSTANDABLE TO LAY COMMITTEE MEMBERS.

SEE NOTES IN STATUS BAR FOR ADVICE ON COMPLETING EACH FIELD. YOU SHOULD READ THE ETHICS APPLICATION GUIDELINES AND HAVE THEM AVAILABLE AS YOU COMPLETE THIS FORM.

APPLICATION FORM

SECTION A APPLICATION DETAILS

A1

Project Title: Coping with crisis - disabled people in emergencies in urban areas

Date of Submission:	Proposed Start Date: 01/09/2015
UCL Ethics Project ID Number: 7111/001	Proposed End Date: 01/09/2016

If this is an application for classroom research as distinct from independent study courses, please provide the following additional details:

Course Title: N/A Course Number: N/A

A2

Principal Researcher

Please note that a student – undergraduate, postgraduate or research postgraduate cannot be the Principal Researcher for Ethics purposes.

Full Name: [redacted]	Position Held: Principal Research Associate
Address: 301 Chadwick Building, CEGE, UCL, WC1E 6BT	Email: [redacted]
	Telephone: [redacted]
	Fax:

Declaration To be Signed by the Principal Researcher

- I have met with and advised the student on the ethical aspects of this project design *(applicable only if the Principal Researcher is not also the Applicant).*
- I understand that it is a UCL requirement for both students & staff researchers to undergo Disclosure and Barring Service (DBS) Checks when working in controlled or regulated activity with children, young people or vulnerable adults. The required DBS Check Disclosure Number(s) is:
- I have obtained approval from the UCL Data Protection Officer stating that the research project is compliant with the Data Protection Act 1998. My Data Protection Registration Number is:
- I am satisfied that the research complies with current professional, departmental and university guidelines including UCL's Risk Assessment Procedures and insurance arrangements.
- I undertake to complete and submit the 'Continuing Review Approval Form' on an annual basis to the UCL Research Ethics Committee.
- I will ensure that changes in approved research protocols are reported promptly and are not initiated without approval by the UCL Research Ethics Committee, except when necessary to eliminate apparent immediate hazards to the participant.
- I will ensure that all adverse or unforeseen problems arising from the research project are reported in a timely fashion to the UCL Research Ethics Committee.
- I will undertake to provide notification when the study is complete and if it fails to start or is abandoned.

SIGNATURE: **DATE:**

A3 **Applicant(s) Details** *(if Applicant is not the Principal Researcher e.g. student details):*

Full Name: Thanos Bantis	
Position Held: EngD student	
Address: [redacted]	Email: [redacted] Telephone: [redacted] Fax:
Full Name: [redacted]	
Position Held: Lecturer	
Address: GM06, Chadwick Building, CEGE, UCL, WC1E 6BT	Email: [redacted] Telephone: [redacted] Fax:

A4 **Sponsor/ Other Organisations Involved and Funding**

a) Sponsor: ☐ **UCL** ☒ **Other institution**
If your project is sponsored by an institution other than UCL please provide details: This project is sponsored by an ESPRC studenship fund

b) Other Organisations: If your study involves another organisation, please provide details. *Evidence that the relevant authority has given permission should be attached or confirmation provided that this will be available upon request.*

c) Funding: What are the sources of funding for this study and will the study result in financial payment or payment in kind to the department or College? *If study is funded solely by UCL this should be stated, the section should not be left blank.* This project is funded through the ESPRC studenship. The research will not result in any form of financial payment to the department or College.

A5 **Signature of Head of Department or Chair of the Departmental Ethics Committee**
(This must not be the same signature as the Principal Researcher)

I have discussed this project with the principal researcher who is suitably qualified to carry out this research and I approve it. The project is registered with the UCL Data Protection Officer, a formal signed risk assessment form has been completed, and appropriate insurance arrangements are in place. *Links to details of UCL's policies on data protection, risk assessment, and insurance arrangements can be found at: http://ethics.grad.ucl.ac.uk/procedures.php*

UCL is required by law to ensure that researchers undergo a Disclosure and Barring Service (DBS) Check if their research project puts them in a position of trust with children under 18 or vulnerable adults.

HEAD OF DEPARTMENT TO DELETE BELOW AS APPLICABLE

I am satisfied that checks: (1) have been satisfactorily completed

If checks are **not** required please clarify why below.

Chair's Action Recommended: ☒ **Yes** ☐ **No**

A recommendation for Chair's action can be based only on the criteria of minimal risk as defined in the Terms of Reference of the UCL Research Ethics Committee.

PRINT NAME: [redacted]

SIGNATURE: **DATE:** 11.06.2015

SECTION B DETAILS OF THE PROJECT

B1 **Please provide a brief summary of the project in simple prose outlining the intended value of the project, giving necessary scientific background** *(max 500 words)*.

A fundamental aspect of individuals' well-being is the ability to reach and engage with the opportunities one values. A person's mobility, although an important aspect towards this goal, it is not sufficient on its own. Attempting to understand people's ability to reach and engage with their day-to-day activities requires a more holistic approach as there is a plethora of other dimensions that influence the degree at which opportunities become accessible. These dimensions can take the form of environmental and socio-economic factors, available material resources and so on.

Moreover, in a dynamic urban environment, disruptions caused by an emergency can also influence an individual's ability to reach and complete their day-to-day activities. Usually, the word "emergency" is used to describe large scale disruptions to a system (such as natural or man-made disasters), however, even small scale perturbations, such as localised public transport service cancellations, can have significant impact to individuals' well-being. This impact is not the same for every individual, with population groups such as mobility impaired people being affected disproportionally. This is especially true given the increased reliance of mobility impaired people on public transport, as well as the transportation disadvantages they face compared to the rest of the population. Even in small localised disruptions, such as limited access to public transport or disrupted bus route services caused by a flood event, there exists evidence of negative impact to mobility impaired people's access to goods and services.

Acknowledging the above, a question naturally rises: How do people with mobility impairments negotiate with such disruptions? Taking the question further: to what extent is peoples' well-being degraded by limited access to goods and services in case of an emergency?

This research aims at being an intersection between accessibility and disability studies, approached from a people's resilience to emergencies viewpoint. Although data related to journey patterns as well as level of disability are regularly collected by state agencies (such as the Office of National Surveys, Department of Transport, Transport For London etc.) and in general allow for structural forms to manifest, they fail to provide detailed information on the individual level. This implication is of great importance in the case of understanding the transportation patterns of mobility impaired people in emergencies, since a disaggregated approach is necessary to observe the amount of variation not only within mobility impaired people (e.g. wheelchair users, crutches users) but also relative to the rest of the population.

To this end, quantitative data collected by a mobile phone application coupled with qualitative questionnaires regarding the users' travel habits, provide the potential to evaluate the influence of different factors in the overall ability of people to cope with potential disturbances to their mobility.

B2 **Briefly characterise in simple prose the research protocol, type of procedure and/or research methodology (e.g. observational, survey research, experimental). Give details of any samples or measurements to be taken** *(max 500 words).*

The part of the research requiring ethical approval is the use of a mobile phone application. This application will log the location of the user at regular time intervals while incorporating a "personal information" and a "travel diary" questionnaires where a user will be able to enter information regarding his/her journeys. This will allow the determination of mobility characteristics for people with different mobility capabilities, and evaluate the significance of different barriers in their ability to complete a set of everyday functions (such as go to work, visit friends etc.)

Specifically, in the personal information section of the application the user will be able to provide information on age, gender, marital status, whether he/she has a disability status, whether he/she uses a mobility aid, a personal assistant, a car and finally the user's occupation status (see attached questionnaires).

The "travel diary" section of the application is designed to give the user the opportunity to specify the purpose of travel, whether he/she is traveling alone, the travel means used and weather these were the preferred choices, as well as the opportunity to report anything that has caused disruptions to the journey. The user will receive a notification at the end of each day, prompting the completion of the questionnaires.

	The time-stamped location logging is done passively, while the application is executed in the background. The accuracy is dependent on the availability of sensors (mobile cell towers, Wi-Fi hotspots and GPS) and can be as coarse as 500m to as accurate as 10m. This will result in a trajectory which the user can view and add qualitative characteristics by means of the travel diary questionnaire. Besides longitude and latitude, the application logs: date, time, accuracy of the estimated location and acceleration. Besides the above, an important variable as recognised in the literature which allows individuals to cope in emergencies is the extent of social networks/social activity. In order to estimate the extent at which the user is socially active, the call details of the user are logged after they are anonymised by the application. Specifically, outgoing, incoming and missed calls are anonymised and logged along with date, time and call durations (if any). Moreover, the application passively scans the environment for Bluetooth devices in the vicinity and logs the corresponding available device IDs. This was done to estimate whether a user is socially active by means of exploiting any regularities in the appearance of Bluetooth devices. All information is stored in a text file in the internal memory of the users' phone which they can then choose to upload to a secure FTP (File Transfer Protocol) server located at UCL computer science department. All information is anonymised by a unique user ID during data saving and uploading. Sensitive information such as phone numbers and Bluetooth device IDs, are encrypted before storage using a hashed-based algorithm. This algorithm obscures the real data, replacing them by a coherent encrypted code. This way, the researcher does know that there where calls made/received to a specific person, but has no way of determining what the phone number of that person is. The same is true for the Bluetooth devices. This high detailed dataset will act as ground truth which can then be used to construct unique spatiotemporal distributions of users' activities depending on the level of mobility impairment. Screenshots of the application interfaces as well as the questionnaires are attached. *Attach any questionnaires, psychological tests, etc. (a standardised questionnaire does not need to be attached, but please provide the name and details of the questionnaire together with a published reference to its prior usage).*

B3	**Where will the study take place (please provide name of institution/department)?** If the study is to be carried out overseas, what steps have been taken to secure research and ethical permission in the study country? Is the research compliant with Data Protection legislation in the country concerned or is it compliant with the UK Data Protection Act 1998? The study will take place in London, UK

B4	**Have collaborating departments whose resources will be needed been informed and agreed to participate?** *Attach any relevant correspondence.* N/A

B5	**How will the results be disseminated, including communication of results with research participants?** All information will be available to the users by means of communicating the student researcher or by means of a dedicated website, where people will have the opportunity to view their trajectories and all other information collected using their unique user ID. The final aggregated results of the overall project will be available to the participants on demand.

B6 **Please outline any ethical issues that might arise from the proposed study and how they are be addressed.** *Please note that all research projects have some ethical considerations so do not leave this section blank.*

Since the focus of the study is evaluating the mobility patterns of population groups such as mobility impaired users, special care has been given so that the application is as non-intrusive as possible, both in the application interface and in the context of the questions within the scope of the project. This was achieved by designing the app such that all information is recorded with minimal user effort/input.

To preserve user anonymity the application doesn't log any personal information revealing the identity of the user such as name, address, telephone number, email etc. Instead, it assigns an identifier code made from randomly collected serial numbers of phones internal sensors (this being the unique user ID). This way, a unique identifier for each phone can be obtained without pointing to the true user's identity.

However, since the application is location aware, in theory a user could be uniquely identified by his/her location. In order to minimise any possibility of this, the data uploading process is done by using a password protected UCL internal server, such that no third party software or services act as mediums (for instance Dropbox, Google Drive etc.).

Another type of sensitive information collected by the application is participants call history and available Bluetooth devices. In order to protect their confidentiality a Secure Hash Algorithm (SHA-1) was used to convert all telephone numbers to unique codes obscuring any links to the original data. Such algorithms are commonly used for password protection purposes.

Another ethical consideration arising from employing a mobile application for data collection purposes is informed consent. Although the users are presented with a description of the type of data collected when they first install the application as well as information on how to withdraw from participating as a dedicated feature within the application (see attachments: "Project Information"), they might have difficulties realising what is the actual goal of the project. To address this issue, they will be given a printed version of the same informed consent for signing when introduced to the application as well as a full debrief of the project in person.

All collected data will be used for the sole purpose of the project and will not be disclosed to third parties.

SECTION C DETAILS OF PARTICIPANTS

C1 **Participants to be studied**

C1a. Number of volunteers:	20-30
Upper age limit:	None
Lower age limit:	18

C1b. Please justify the age range and sample size:

The age range is reflecting the need to include most age groups in a position to use the public transport for their accessibility needs. The sample size is reflecting the need to include a sufficient sample for statistical analysis.

C2 **If you are using data or information held by a third party, please explain how you will obtain this. You should confirm that the information has been obtained in accordance with the UK Data Protection Act 1998.**

N/A

C3

Will the research include children or vulnerable adults such as individuals with a learning disability or cognitive impairment or individuals in a dependent or unequal relationship? ☐ **Yes** ☒ **No**

How will you ensure that participants in these groups are competent to give consent to take part in this study? *If you have relevant correspondence, please attach it.*

Although the nature of the data collection method could potentially involve people with severe mobility impairments in need of a carer or personal assistant, it is not within the scope of the project to include people that do not personally consent in the use of the data generated by the app for research purposes. In this view, the informed consent and information sheet presented within the lifecycle of the app is intending to provide absolute transparency to the user on the way the data are collected and treated.

C4

Will payment or any other incentive, such as gift service or free services, be made to any research participant?

☒ **Yes** ☐ **No**

If yes, please specify the level of payment to be made and/or the source of the funds/gift/free service to be used.

Although this is not fully specified at this stage, most likely amazon vouchers will be provided as an incentive to participating. Since there are no other monetary sources for funding, this will be covered by part of the student's available research funds.

Please justify the payment/other incentive you intend to offer.

C5

Recruitment

(i) Describe how potential participants will be identified:

Self-identified

(ii) Describe how potential participants will be approached:

Primarily through UCL's Transport Accessibility Rehabilitation Services Advisory Network webpage (http://www.cege.ucl.ac.uk/tarsan/Pages/TARSANEvents.aspx)

(iii) Describe how participants will be recruited:

After contacting the student researcher they will be given details on downloading, installing and using the app.

Attach recruitment emails/adverts/webpages. A data protection disclaimer should be included in the text of such literature.

C6

Will the participants participate on a fully voluntary basis? ☒ **Yes** ☐ **No**

Will UCL students be involved as participants in the research project? ☒ **Yes** ☐ **No**

If yes, care must be taken to ensure that they are recruited in such a way that they do not feel any obligation to a teacher or member of staff to participate.

Please state how you will bring to the attention of the participants their right to withdraw from the study without penalty?

The Project information section of the app will specify ways that people can withdraw from the research should they wish to. UCL students could participate on a fully voluntary basis after being self-identified.

C7 **CONSENT**

Please describe the process you will use when seeking and obtaining consent.

Consent will be obtained by using the lifecycle of the application. This is described as follows:

When the user installs the application, the application will navigate to an Information Consent form where information regarding the objective, aim, potential benefits and data gathered will be provided. The user will then be given the option to agree and continue to the main application, or disagree. In the second case, the application will automatically exit. Once the user chooses to agree, this initial disclaimer will not be presented again to avoid redundancy. Additional information about ways to communicate, withdraw the data or any other queries will be given in the main menu of the application. Please see attached screenshots for more information.

Furthermore, a full debriefing will be given in person during the recruitment process as well as physical copies for Informed Consent and Information Sheet.

A copy of the participant information sheet and consent form must be attached to this application. For your convenience proformas are provided in C10 below. These should be filled in and modified as necessary.

In cases where it is not proposed to obtain the participants informed consent, please explain why below

C8 **Will any form of deception be used that raises ethical issues? If so, please explain.**

N/A

C9 **Will you provide a full debriefing at the end of the data collection phase?** ☐ **Yes** ☒ **No**

If 'No', please explain why below.

User will be offered the option to retrieve his data at any point. More information about the uses and types of data collected will be integrated in the lifecycle of the application.

C10 **Information Sheets And Consent Forms**

Please see attached.

A poorly written Information Sheet(s) and Consent Form(s) that lack clarity and simplicity frequently delay ethics approval of research projects. The wording and content of the Information Sheet and Consent Form must be appropriate to the age and educational level of the research participants and clearly state in simple non-technical language what the participant is agreeing to. Use the active voice e.g. "we will book" rather than "bookings will be made". Refer to participants as "you" and yourself as "I" or "we". An appropriate translation of the Forms should be provided where the first language of the participants is not English. If you have different participant groups you should provide Information Sheets and Consent Forms as appropriate (e.g. one for children and one for parents/guardians) using the templates below. Where children are of a reading age, a written Information Sheet should be provided. When participants cannot read or the use of forms would be inappropriate, a description of the verbal information to be provided should be given. Please ensure that you trial the forms on an age-appropriate person before you submit your application.

SECTION D DETAILS OF RISKS AND BENEFITS TO THE RESEARCHER AND THE RESEARCHED

D1 **Have UCL's Risk Assessment Procedures been followed?** ☐ **Yes** ☒ **No**

If **No**, please explain.

The project will not deal with any hazardous processes or materials. The application is designed to be non-intrusive, requiring minimal physical or mental interaction to participants, hence no ergonomic hazards due to repetition, awkward postures etc. were identified.

D2 **Does UCL's insurer need to be notified about your project before insurance cover can be provided?** ☐ **Yes** ☒ **No**

The insurance for all UCL studies is provided by a commercial insurer. For the majority of studies the cover is automatic. However, for a minority of studies, in certain categories, the insurer requires prior notification of the project before cover can be provided.

If **Yes**, please provide confirmation that the appropriate insurance cover has been agreed. *Please attach your UCL insurance registration form and any related correspondence.*

N/A

D3 **Please state briefly any precautions being taken to protect the health and safety of researchers and others associated with the project (as distinct from the research participants).**

N/A

D4 **Will these participants participate in any activities that may be potentially stressful or harmful in connection with this research?** ☐ **Yes** ☒ **No**

If **Yes**, please describe the nature of the risk or stress and how you will minimise and monitor it.

N/A

D5 **Will group or individual interviews/questionnaires raise any topics or issues that might be sensitive, embarrassing or upsetting for participants?**

☐ **Yes** ☒ **No**

If **Yes,** please explain how you will deal with this.

N/A

D6

Please describe any expected benefits to the participant.

Although no direct benefits to the applicant have been identified, the users will be given the opportunity to view and download their trajectory data as well as some summary statistics of their activity. This will be done via a dedicated website.

D7

Specify whether the following procedures are involved:

Any invasive procedure(s) ☐ **Yes** ☒ **No**

Physical contact ☐ **Yes** ☒ **No**

Any procedure(s) that may cause mental distress ☐ **Yes** ☒ **No**

Please state briefly any precautions being taken to protect the health and safety of the research participants.

N/A

D8

Does the research involve the use of drugs? ☐ **Yes** ☒ **No**

If **Yes**, please name the drug/product and its intended use in the research and then complete Appendix I

N/A

Does the project involve the use of genetically modified materials? ☐ **Yes** ☒ **No**

If **Yes**, has approval from the Genetic Modification Safety Committee been obtained for work? ☐ Yes ☐ No

If **Yes**, please quote the Genetic Modification Reference Number:

D9

Will any non-ionising radiation be used on the research participant(s)? ☐ **Yes** ☒ **No**

If **Yes**, please complete Appendix II.

D10

Are you using a medical device in the UK that is CE-marked and is being used within its product indication? ☐ **Yes** ☒ **No**

If **Yes**, please complete Appendix III.

CHECKLIST

Please submit ether 12 copies (1 original + 11 double sided photocopies) of your completed application form for full committee review or 3 copies (1 original + 2 double sided copies) for chair's action, together with the appropriate supporting documentation from the list below to the UCL Research Ethics Committee Administrator. You should also submit your application form electronically to the Administrator at: ethics@ucl.ac.uk

Documents to be Attached to Application Form (if applicable)	Ticked if attached	Tick if not relevant
Section B: Details of the Project		
• Questionnaire(s) / Psychological Tests	☒	☐
• Relevant correspondence relating to involvement of collaborating department/s and agreed participation in the research.	☐	☒
Section C: Details of Participants		
• Parental/guardian consent form for research involving participants under 18	☐	☒
• Participant/s information sheet	☒	☐
• Participant/s consent form/s	☒	☐
• Advertisement	☐	☒
Section D: Details of Risks and Benefits to the Researcher and the Researched		
• Insurance registration form and related correspondence	☐	☒
Appendix I: Research Involving the Use of Drugs		
• Relevant correspondence relating to agreed arrangements for dispensing ☒ with the pharmacy	☐	
• Written confirmation from the manufacturer that the drug/substance has has been manufactured to GMP	☐	☒
• Proposed volunteer contract	☐	☒
• Full declaration of financial or direct interest	☐	☒
• Copies of certificates: CTA etc…	☐	☒
Appendix II: Use of Non-Ionising Radiation		
Appendix III: Use Medical Devices		

Please note that correspondence regarding the application will normally be sent to the Principal Researcher and copied to other named individuals.

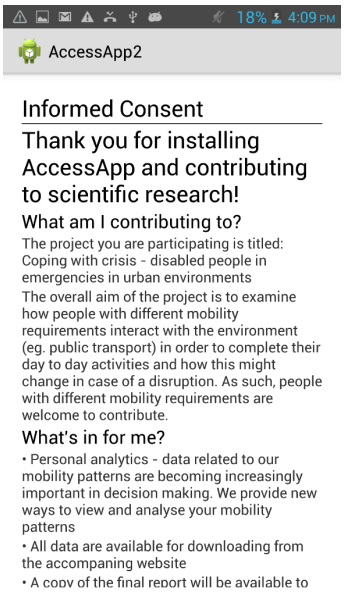
Questionnaires:

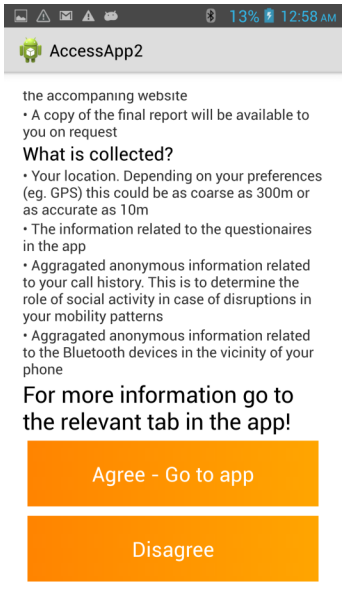
Personal Information		**Travel Diary**	
Age	18-22	**What was the purpose of today's travel? (Choose more than one journey if necessary)**	Work
	23-39		Education
	40-59		Medical
	60+		Shopping
Gender	Male		Visit friends / Family
	Female		Leisure / Entertainment
Marital Status	Single	**Who were you travelling with?**	Travelling alone
	Married / Living with partner / Living with family		Personal Assistant
Registered disabled	Yes		Friends
	No		Family
			Work / Business colleagues

Mobility Aid	No mobility requirements	**What travel means did you choose? (Choose more than one if necessary)**	Walk
	Wheelchair (self-propelled)		Tube
	Wheelchair (attendant-propelled)		National Rail / Overground
	Mobility scooter		Car (as a driver)
	Crutches		Car (as a passenger)
	Other		Taxi / Minicab
Occupation	Employed full time		Didn't travel
	Employed part-time		Other (please specify in the textbox below)
	Unemployed	**Was the chosen travel means your first choice?**	Yes
	Student		No
	Retired	**Did you experience anything unusual that disrupted your trip?**	Feeling unwell
Do you have a Personal Assistant?	All the time		Bad weather

	Some of the time		Disrupted service
	Rarely		Social discrimination
	Never		Other (please specify in the textbox below)
Do you have access to car?	All the time		
	Some of the time		
	Rarely		
	Never		

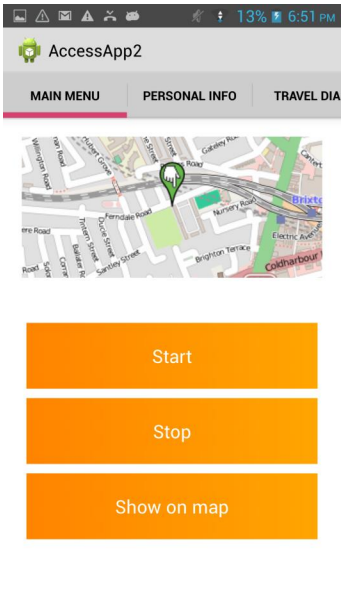
1) Informed Consent View, appearing when the user first installs the app





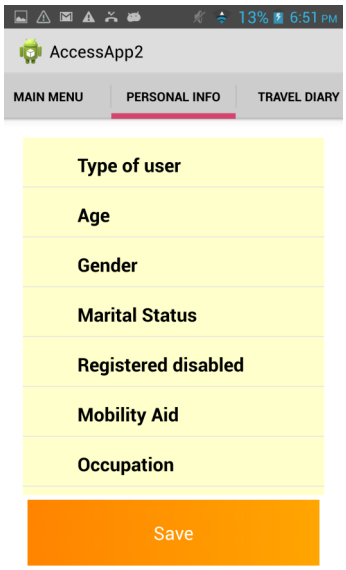
Text:
Informed Consent: Thank you for installing AccessApp and contributing to scientific research! What am I contributing to? The project you are participating is titled: Coping with crisis - disabled people in emergencies in urban environments The overall aim of the project is to examine how people with different mobility requirements interact with the environment (eg. public transport) in order to complete their day to day activities and how this might change in case of a disruption. As such, people with different mobility requirements are welcome to contribute. What's in for me? Personal analytics - data related to our mobility patterns are becoming increasingly important in decision making. We provide new ways to view and analyse your mobility patterns All data are available for downloading from the accompanying website
A copy of the final report will be available to you on request
What is collected?
Your location. Depending on your preferences (e.g. GPS) this could be as coarse as 300m or as accurate as 10m
The information related to the questionnaires in the app
Aggregated, anonymous information related to your call history. This is to determine the role of social activity in case of disruptions in your mobility patterns.
Aggregated, anonymous information related to the Bluetooth devices in the vicinity of your phone.
For more information go to the relevant tab in the app!

2) Main Menu. Appears when user taps on Agree – Go to app button



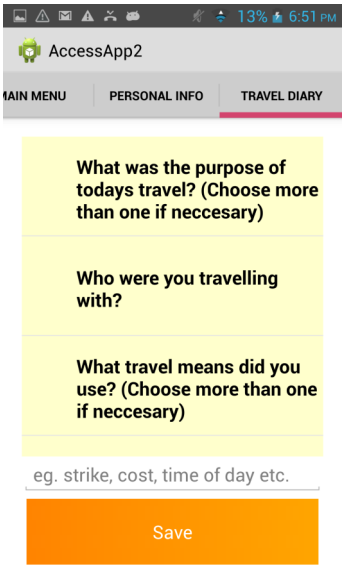
Start Button: App starts logging all relevant information passively. Specifically: Time stamped network based location, bearing, heading, acceleration, call history, Bluetooth devices
Stop Button: App stops and all data are uploaded on the server
Show on map Button: Shows the past locations of the user

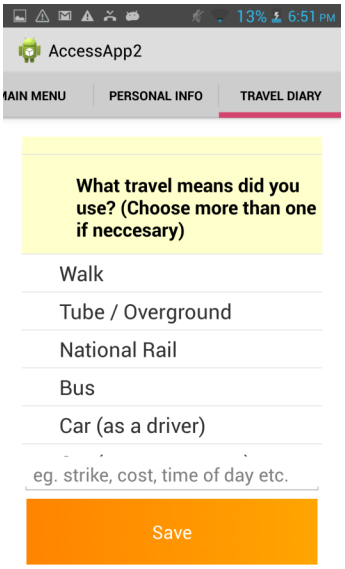
3) Personal Info



Save Button: Saves and uploads all information to the server

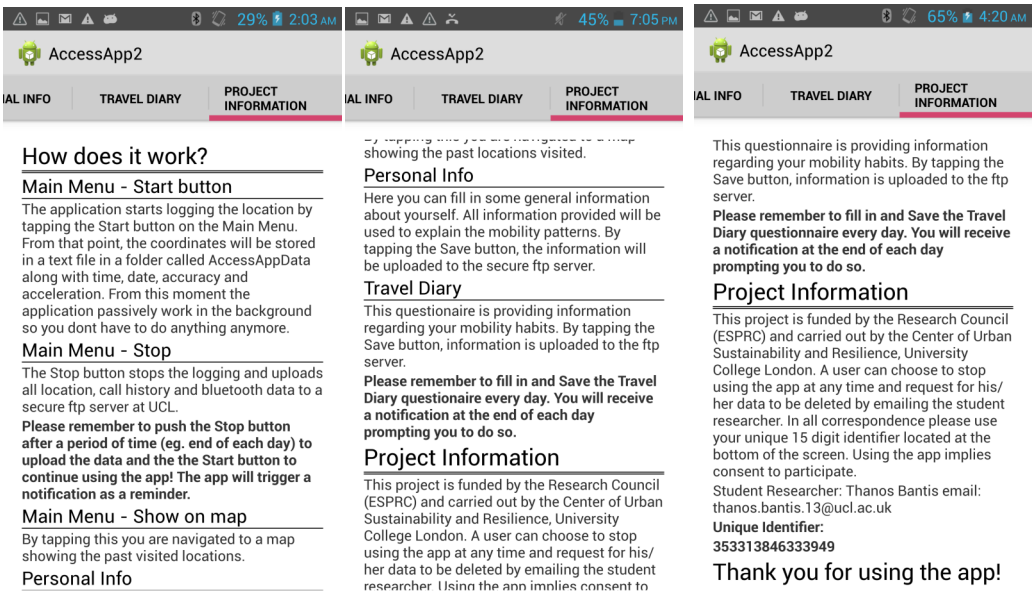
4) Travel Diary





Save Button: Saves and uploads all information to the server

5) Project Information



Text: How does it work? Main Menu - Start button
The application starts logging the location by tapping the Start button on the Main Menu. From that point, the coordinates will be stored in a text file in a folder called AccessAppData along with time, date, accuracy and acceleration.
From this moment the application passively work in the background so you don't have to do anything anymore.
Main Menu - Stop
The Stop button stops the logging and uploads all location, call history and Bluetooth data to a secure ftp server at UCL.
Please remember to push the Stop button after a period of time (e.g. end of each day) to upload the data and the Start button to continue using the app! The app will trigger a notification as a reminder.
Main Menu - Show on map
By tapping this you are navigated to a map showing the past visited locations.

Personal Info
Here you can fill in some general information about yourself. All information provided will be used to explain the mobility patterns. By tapping the Save button, the information will be uploaded to the secure ftp server.
Travel Diary
This questionnaire is providing information regarding your mobility habits. By tapping the Save button, information is uploaded to the ftp server.
Please remember to fill in and Save the Travel Diary questionnaire every day. You will receive a notification at the end of each day prompting you to do so.
Project Information
This project is funded by the Research Council (ESPRC) and carried out by the Centre of Urban Sustainability and Resilience, University College London. A user can choose to stop using the app at any time and request for his/her data to be deleted by emailing the student researcher. Using the app implies consent to participate.
Student Researcher: Thanos Bantis email: thanos.bantis.13@ucl.ac.uk
Thank you for using the app!

Forms

Information Sheet for in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Coping with crisis - disabled people in emergencies in urban areas**

This study has been approved by the UCL Research Ethics Committee (Project ID Number): **7111/001**

Name

Work Address

Contact Details (*For students, we strongly advise against the use of a personal contact number)

We would like to invite to participate in this research project.

Details of Study:

Thank you for installing AccessApp and contributing to scientific research!

What am I contributing to?

The project you are participating is titled: Coping with crisis - disabled people in emergencies in urban environments. The overall aim of the project is to examine how people with different mobility requirements interact with the environment (e.g. public transport) in order to complete their day to day activities and how this might change in case of a disruption. As such, people with different mobility requirements are welcome to contribute.

What's in for me?

- Personal analytics - data related to our mobility patterns are becoming increasingly important in decision making. We provide new ways to view and analyse your mobility patterns. All data are available for downloading from the accompanying website, or by contacting the student researcher by email. Since all your data is automatically anonymised please use the unique identifier located at the bottom of Project Information tab
- A copy of the final report will be available to you on request

What is collected?

- Your location. Depending on your preferences (e.g. GPS) this could be as coarse as 300m or as accurate as 10m
- The information related to the questionnaires in the app
- Aggregated, anonymous information related to your call history. This is to determine the role of social activity in case of disruptions in your mobility patterns.
- Aggregated, anonymous information related to the Bluetooth devices in the vicinity of your phone.

How does it work?

The research is conducted by the use of AccessApp mobile phone app. The use of the app is very simple, below is a quick walkthrough the functionalities and what each button does.

Main Menu - Start button

The application starts logging the location by tapping the Start button on the Main Menu. From that point, the coordinates will be stored in a text file in a folder called AccessAppData along with time, date, accuracy and acceleration.

From this moment the application passively work in the background so you don't have to do anything anymore.

Main Menu - Stop button

The Stop button stops the logging and uploads all location, call history and Bluetooth data to a secure ftp server at UCL. **Please remember to push the Stop button after a period of time (e.g. end of each day) to upload the data and the Start button to continue using the app! The app will trigger a notification as a reminder.**

Main Menu - Show on map button

By tapping this you are navigated to a map showing the past visited locations.

Personal Info

Here you can fill in some general information about yourself. All information provided will be used to explain the mobility patterns. By tapping the Save button, the information will be uploaded to the secure ftp server.

Travel Diary

This questionnaire is providing information regarding your mobility habits. By tapping the Save button, information is uploaded to the ftp server.

Please remember to fill in and Save the Travel Diary questionnaire every day. You will receive a notification at the end of each day prompting you to do so.

Project Information

This project is funded by the Research Council (ESPRC) and carried out by the Centre of Urban Sustainability and Resilience, University College London. A user can choose to stop using the app at any time and request for his/her data to be deleted by emailing the student researcher. In all correspondence please use your unique 15 digit identifier located at the bottom of the screen. Using the app implies consent to participate.

Student Researcher: Thanos Bantis email: thanos.bantis.13@ucl.ac.uk

Please discuss the information above with others if you wish or ask us if there is anything that is not clear or if you would like more information.

It is up to you to decide whether to take part or not; choosing not to take part will not disadvantage you in any way. If you do decide to take part you are still free to withdraw at any time and without giving a reason.

All data will be collected and stored in accordance with the Data Protection Act 1998.Thank you for reading this information sheet and for considering take part in this research.

Informed Consent Form for in Research Studies

Please complete this form after you have read the Information Sheet and/or listened to an explanation about the research.

Title of Project: **Coping with crisis - disabled people in emergencies in urban areas**

This study has been approved by the UCL Research Ethics Committee (Project ID Number): **7111/001**

Thank you for your interest in taking part in this research. Before you agree to take part, the person organising the research must explain the project to you.

If you have any questions arising from the Information Sheet or explanation already given to you, please ask the researcher before you to decide whether to join in. You will be given a copy of this Consent Form to keep and refer to at any time.

Participant's Statement

I

- have read the notes written above and the Information Sheet, and understand what the study involves.
- understand the nature of data gathered by the use of the app.
- understand that if I decide at any time that I no longer wish to take part in this project, I can uninstall the app. If I still wish to collect my data I can contact the student researcher by using the unique Identifier located at the bottom of the Project Information tab in the app
- understand that the information I have submitted will be published as a report and I will be sent a copy if I wish to. Confidentiality and anonymity will be maintained and it will not be possible to identify me from any publications.
- consent to the processing of my personal information for the purposes of this research study.
- understand that such information will be treated as strictly confidential and handled in accordance with the provisions of the Data Protection Act 1998.
- agree that the research project named above has been explained to me to my satisfaction and I agree to take part in this study.

Signed: Date:

Bibliography

Ahlfeldt, G. (2011), 'If alonso was right: modeling accessibility and explaining the residential land gradient', *Journal of Regional Science* **51**(2), 318–338.

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.

Alkire, S. (2008), 'Using the capability approach: Prospective and evaluative analyses", *The Capability Approach: Concepts, Measures and Applications, Cambridge University Press, Cambridge* pp. 26–50.

Allahviranloo, M. & Recker, W. (2013), 'Daily activity pattern recognition by using support vector machines with multiple classes', *Transportation Research Part B: Methodological* **58**, 16–43.

Allahviranloo, M. & Recker, W. (2015), 'Mining activity pattern trajectories and allocating activities in the network', *Transportation* **42**(4), 561–579.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L. & Hickman, M. (2018), 'Public transport trip purpose inference using smart card fare data', *Transportation Research Part C: Emerging Technologies* **87**, 123–137.

Anand, P., Krishnakumar, J. & Tran, N. B. (2011), 'Measuring welfare: Latent variable models for happiness and capabilities in the presence of unobservable heterogeneity', *Journal of public economics* **95**(3-4), 205–215.

Anda, C., Erath, A. & Fourie, P. J. (2017), 'Transport modelling in the age of big data', *International Journal of Urban Sciences* **21**(sup1), 19–42.

Anselin, L. (2013), *Spatial econometrics: methods and models*, Vol. 4, Springer Science & Business Media.

Arcidiacono, P. & Miller, R. A. (2011), 'Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity', *Econometrica* **79**(6), 1823–1867.

Bagchi, M. & White, P. R. (2005), 'The potential of public transport smart card data', *Transport Policy* **12**(5), 464–474.

Balke, A. & Pearl, J. (1994), Counterfactual probabilities: Computational methods, bounds and applications, *in* 'Uncertainty Proceedings 1994', Elsevier, pp. 46–54.

Bantis, T. & Haworth, J. (2017), 'Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics', *Transportation Research Part C: Emerging Technologies* **80**, 286–309.

Bantis, T. & Haworth, J. (2019), 'Non-employment activity type imputation from points of interest and mobility data at an individual level: How accurate can we get?', *ISPRS International Journal of Geo-Information* **8**(12), 560.

Bantis, T. & Haworth, J. (2020), 'Assessing transport related social exclusion using a capabilities approach to accessibility framework: A dynamic bayesian network approach', *Journal of Transport Geography* **84**(102673).

Bantis, T., Haworth, J., Holloway, C. & Twigg, J. (2015), Mapping spatio-temporal patterns of disabled people in emergencies: A bayesian approach, *in* 'Proceedings of GeoComputation 2015 Conference'.

Bantis, T., Haworth, J., Holloway, C. & Twigg, J. (2017), Mapping spatiotemporal patterns of disabled people: The case of the st. jude's storm emergency, *in* 'Advances in Geocomputation', Springer, pp. 97–113.

Barredo, J. I., Kasanko, M., McCormick, N. & Lavalle, C. (2003), 'Modelling dynamic spatial processes: simulation of urban future scenarios through cellular automata', *Landscape and urban planning* **64**(3), 145–160.

Barry, J., Freimer, R. & Slavin, H. (2009), 'Use of entry-only automatic fare collection data to estimate linked transit trips in new york city', *Transportation Research Record: Journal of the Transportation Research Board* (2112), 53–61.

Barry, J., Newhouser, R., Rahbee, A. & Sayeda, S. (2002), 'Origin and destination estimation in new york city with automated fare system data', *Transportation Research Record: Journal of the Transportation Research Board* (1817), 183–187.

Batty, M. (2009), 'Accessibility: in search of a unified theory'.

Ben-Akiva, M. (1979), 'Disaggregate travel and mobility choice models and measures of accessibility', *Behavioural travel modelling* .

Ben-Akiva, M. & Bierlaire, M. (1999), Discrete choice methods and their applications to short term travel decisions, *in* 'Handbook of transportation science', Springer, pp. 5–33.

Ben-Akiva, M. & Lerman, S. R. (1985), 'Discrete choice analysis: Theory and application to'.

Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D. S. et al. (2002), 'Hybrid choice models: Progress and challenges', *Marketing Letters* **13**(3), 163–175.

Ben-Elia, E. & Benenson, I. (2019), 'A spatially-explicit method for analyzing the equity of transit commuters' accessibility', *Transportation Research Part A: Policy and Practice* **120**, 31–42.

Bertuccelli, L. F. & How, J. P. (2008), Estimation of non-stationary markov chain transition models, *in* '2008 47th IEEE Conference on Decision and Control', IEEE, pp. 55–60.

Besag, J., York, J. & Mollié, A. (1991), 'Bayesian image restoration, with two applications in spatial statistics', *Annals of the Institute of Statistical Mathematics* **43**(1), 1–20.

Beyazit, E. (2011), 'Evaluating social justice in transport: lessons to be learned from the capability approach', *Transport reviews* **31**(1), 117–134.

Bhandari, K., Kato, H. & Hayashi, Y. (2009), 'Economic and equity evaluation of delhi metro', *International Journal of Urban Sciences* **13**(2), 187–203.

Bhat, C. R. & Guo, J. (2004), 'A mixed spatially correlated logit model: formulation and application to residential choice modeling', *Transportation Research Part B: Methodological* **38**(2), 147–168.

Bills, T. S. & Walker, J. L. (2017), 'Looking beyond the mean for equity analysis: Examining distributional impacts of transportation improvements', *Transport Policy* **54**, 61–69.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Blythe, P. T. (2004), Improving public transport ticketing through smart cards, *in* 'Proceedings of the Institution of Civil Engineers-Municipal Engineer', Vol. 157, Thomas Telford Ltd, pp. 47–54.

Bocarejo S, J. P. & Oviedo H, D. R. (2012), 'Transport accessibility and social inequities: a tool for identification of mobility needs and evaluation of transport investments', *Journal of Transport Geography* **24**, 142–154.

Bohte, W. & Maat, K. (2009), 'Deriving and validating trip purposes and travel modes for multi-day gps-based travel surveys: A large-scale application in the netherlands', *Transportation Research Part C: Emerging Technologies* **17**(3), 285–297.

Bolbol, A., Cheng, T., Tsapakis, I. & Haworth, J. (2012), 'Inferring hybrid transportation modes from sparse gps data using a moving window svm classification', *Computers, Environment and Urban Systems* **36**(6), 526–537.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P. & Snelson, E. (2013), 'Counterfactual reasoning and learning systems: The example of computational advertising', *The Journal of Machine Learning Research* **14**(1), 3207–3260.

Bradley, A. P. (1997), 'The use of the area under the roc curve in the evaluation of machine learning algorithms', *Pattern recognition* **30**(7), 1145–1159.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L. et al. (2015), 'Inferring causal impact using bayesian structural time-series models', *The Annals of Applied Statistics* **9**(1), 247–274.

Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L. (2011), *Handbook of markov chain monte carlo*, Chapman and Hall/CRC.

Browne, W. J. (2004), 'An illustration of the use of reparameterisation methods for improving mcmc efficiency in crossed random effect models', *Multilevel modelling newsletter* **16**(1), 13–25.

Brownstone, D. (2001), 'Discrete choice modeling for transportation'.

Buntine, W. L. (1995), Chain graphs for learning, *in* 'Proceedings of the Eleventh conference on Uncertainty in artificial intelligence', Morgan Kaufmann Publishers Inc., pp. 46–54.

Burns, L. D. (1980), 'Transportation, temporal, and spatial components of accessibility'.

Camporeale, R., Caggiani, L., Fonzone, A. & Ottomanelli, M. (2017), 'Quantifying the impacts of horizontal and vertical equity in transit route planning', *Transportation Planning and Technology* **40**(1), 28–44.

Cao, M. & Hickman, R. (2019*a*), 'Understanding travel and differential capabilities and functionings in beijing', *Transport policy* **83**, 46–56.

Cao, M. & Hickman, R. (2019*b*), 'Urban transport and social inequities in neighbourhoods near underground stations in greater london', *Transportation planning and technology* **42**(5), 419–441.

Casas, I. (2007), 'Social exclusion and the disabled: An accessibility approach', *The Professional Geographer* **59**(4), 463–477.

Casas, I., Horner, M. W. & Weber, J. (2009), 'A comparison of three methods for identifying transport-based exclusion: a case study of children's access to urban opportunities in erie and niagara counties, new york', *International Journal of Sustainable Transportation* **3**(4), 227–245.

Cepolina, E. M. & Tyler, N. (2004), 'Microscopic simulation of pedestrians in accessibility evaluation', *Transportation planning and technology* **27**(3), 145–180.

Ceriani, L., Gigliarano, C. et al. (2016), Multidimensional well-being: A bayesian networks approach, Technical report.

Chang, Y. & Zhao-Cheng, H. (2016), 'Travel pattern recognition using smart card data in public transit', *International Journal of Emerging Engineering Research and Technology* **6**.

Chapleau, R., Trépanier, M. & Chu, K. K. (2008), The ultimate survey for transit planning: Complete information with smart card data and gis, *in* 'Proceedings of the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability', pp. 25–31.

Chen, B. Y., Li, Q., Wang, D., Shaw, S.-L., Lam, W. H., Yuan, H. & Fang, Z. (2013), 'Reliable space–time prisms under travel time uncertainty', *Annals of the Association of American Geographers* **103**(6), 1502–1521.

Chen, B. Y., Wang, Y., Wang, D. & Lam, W. H. (2019), 'Understanding travel time uncertainty impacts on the equity of individual accessibility', *Transportation Research Part D: Transport and Environment* **75**, 156–169.

Chen, B. Y., Wang, Y., Wang, D., Li, Q., Lam, W. H. & Shaw, S.-L. (2018), 'Understanding the impacts of human mobility on accessibility using massive mobile phone tracking data', *Annals of the American Association of Geographers* **108**(4), 1115–1133.

Chen, C., Ma, J., Susilo, Y., Liu, Y. & Wang, M. (2016), 'The promises of big data and small data for travel behavior (aka human mobility) analysis', *Transportation Research Part C: Emerging Technologies* **68**, 285–299.

Cheng, J. & Bertolini, L. (2013), 'Measuring urban job accessibility with distance decay, competition and diversity', *Journal of Transport geography* **30**, 100–109.

Chikaraishi, M. (2017), Mobility of the elderly, *in* 'Life-oriented behavioral research for urban policy', Springer, pp. 267–291.

Chikaraishi, M., Jana, A., Bardhan, R., Varghese, V. & Fujiwara, A. (2017), 'A framework to analyze capability and travel in formal and informal urban settings: a case from mumbai', *Journal of transport geography* **65**, 101–110.

Chorus, C. G. & De Jong, G. C. (2011), 'Modeling experienced accessibility for utility-maximizers and regret-minimizers', *Journal of Transport Geography* **19**(6), 1155–1162.

Church, A., Frost, M. & Sullivan, K. (2000), 'Transport and social exclusion in london', *Transport Policy* **7**(3), 195–205.

Church, R. L. & Marston, J. R. (2003), 'Measuring accessibility for people with a disability', *Geographical Analysis* **35**(1), 83–96.

Cohen, J. E. & Kempermann, J. (1998), *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*, Springer Science & Business Media.

Comim, F. (2003), 'Capability dynamics: the importance of time to capability assessments'.

Comim, F. (2008), 'Measuring capabilities', *The capability approach: concepts, measures and application. Cambridge UP, Cambridge* pp. 157–200.

Congdon, P. (2003), *Applied bayesian modelling*, John Wiley & Sons.

Congdon, P. (2007), *Bayesian statistical modelling*, Vol. 704, John Wiley & Sons.

Congdon, P. D. (2010), *Applied Bayesian hierarchical methods*, CRC Press.

Cui, J., Liu, F., Janssens, D., An, S., Wets, G. & Cools, M. (2016), 'Detecting urban road network accessibility problems using taxi gps data', *Journal of Transport Geography* **51**, 147–157.

Currie, G., Richardson, T., Smyth, P., Vella-Brodrick, D., Hine, J., Lucas, K., Stanley, J., Morris, J., Kinnear, R. & Stanley, J. (2010), 'Investigating links between transport disadvantage, social exclusion and well-being in melbourne–updated results', *Research in transportation economics* **29**(1), 287–295.

Daziano, R. A., Miranda-Moreno, L. & Heydari, S. (2013), 'Computational bayesian statistics in transportation modeling: from road safety analysis to discrete choice', *Transport reviews* **33**(5), 570–592.

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**(3), 245–268.

de Jong, R., Pieters, M., Daly, A., Graafland, I., Kroes, E. & Koopmans, C. (2005), Using the logsum as an evaluation measure: Literature and case study, final report, Technical report, WR-275-AVV, Transport Research Centre of the Dutch Ministry of Transport, RAND Europe, Leiden.

de Palma, A. & Kilani, K. (2005), 'Switching in the logit', *Economics Letters* **88**(2), 196–202.

De Palma, A. & Kilani, K. (2011), 'Transition choice probabilities and welfare analysis in additive random utility models', *Economic Theory* **46**(3), 427–454.

Delafontaine, M., Neutens, T., Schwanen, T. & Van de Weghe, N. (2011), 'The impact of opening hours on the equity of individual space–time accessibility', *Computers, environment and urban systems* **35**(4), 276–288.

Delamater, P. L. (2013), 'Spatial accessibility in suboptimally configured health care systems: A modified two-step floating catchment area (m2sfca) metric', *Health & place* **24**, 30–43.

Delbosc, A. & Currie, G. (2011*a*), 'The spatial context of transport disadvantage, social exclusion and well-being', *Journal of Transport Geography* **19**(6), 1130–1137.

Delbosc, A. & Currie, G. (2011*b*), 'Using lorenz curves to assess public transport equity', *Journal of Transport Geography* **19**(6), 1252–1259.

Deneulin, S. (2008), 'Beyond individual freedom and agency: Structures of living together in sen's capability approach to development'.

Department for Transport (2014), 'Accessibility indicators'. Online; Accessed 13/09/2016.

URL: *https://data.gov.uk/dataset/8d72834b-c338-4e03-9da7-7e9016b74a6b/accessibility-statistics*

Devillaine, F., Munizaga, M. & Trépanier, M. (2012), 'Detection of activities of public transport users by analyzing smart card data', *Transportation Research Record: Journal of the Transportation Research Board* (2276), 48–55.

DfT (2014), 'Manual for streets. london: Department for transport'. Online; Accessed 13/01/2018.
URL: *https://www.gov.uk/government/uploads/system/uploads/*

Di Tommaso, M. L. (2007), 'Children capabilities: A structural equation model for india', *The Journal of Socio-Economics* **36**(3), 436–450.

Dodson, J., Buchanan, N., Gleeson, B. & Sipe, N. (2006), 'Investigating the social dimensions of transport disadvantage—i. towards new concepts and methods', *Urban Policy and Research* **24**(4), 433–453.

Doi, K., Kii, M. & Nakanishi, H. (2008), 'An integrated evaluation method of accessibility, quality of life, and social interaction', *Environment and Planning B: Planning and Design* **35**(6), 1098–1116.

Dong, G., Ma, J., Harris, R. & Pryce, G. (2016), 'Spatial random slope multilevel modeling using multivariate conditional autoregressive models: A case study of subjective travel satisfaction in beijing', *Annals of the American Association of Geographers* **106**(1), 19–35.

Dong, X., Ben-Akiva, M. E., Bowman, J. L. & Walker, J. L. (2006), 'Moving from trip-based to activity-based measures of accessibility', *Transportation Research Part A: policy and practice* **40**(2), 163–180.

Duarte, C. W., Klimentidis, Y. C., Harris, J. J., Cardel, M. & Fernández, J. R. (2011), A hybrid bayesian network/structural equation modeling (bn/sem) approach for detecting physiological networks for obesity-related genetic variants, *in* 'Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on', IEEE, pp. 696–702.

Eftekhari, H. R. & Ghatee, M. (2016), 'An inference engine for smartphones to preprocess data and detect stationary and transportation modes', *Transportation Research Part C: Emerging Technologies* **69**, 313–327.

El-Geneidy, A., Buliung, R., Diab, E., van Lierop, D., Langlois, M. & Legrain, A. (2016), 'Non-stop equity: Assessing daily intersections between transit accessi-

bility and social disparity across the greater toronto and hamilton area (gtha)', *Environment and Planning B: Planning and Design* **43**(3), 540–560.

El-Geneidy, A., Levinson, D., Diab, E., Boisjoly, G., Verbich, D. & Loong, C. (2016), 'The cost of equity: Assessing transit accessibility and social disparity using total travel cost', *Transportation Research Part A: Policy and Practice* **91**, 302–316.

Ettema, D. & Timmermans, H. (2007), 'Space–time accessibility under conditions of uncertain travel times: theory and numerical simulations', *Geographical Analysis* **39**(2), 217–240.

Evans, G. (2009), 'Accessibility, urban design and the whole journey environment', *Built environment* **35**(3), 366–385.

Farber, S. & Fu, L. (2017), 'Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis) investments over time', *Computers, Environment and Urban Systems* **62**, 30–40.

Farber, S., Morang, M. Z. & Widener, M. J. (2014), 'Temporal variability in transit-based accessibility to supermarkets', *Applied Geography* **53**, 149–159.

Farrington, J. & Farrington, C. (2005), 'Rural accessibility, social inclusion and social justice: towards conceptualisation', *Journal of Transport geography* **13**(1), 1–12.

Farrington, J. H. (2007), 'The new narrative of accessibility: its potential contribution to discourses in (transport) geography', *Journal of Transport Geography* **15**(5), 319–330.

Fearnley, N., Currie, G., Flügel, S., Gregersen, F. A., Killi, M., Toner, J. & Wardman, M. (2018), 'Competition and substitution between public transport modes', *Research in Transportation Economics* **69**, 51–58.

Feng, T. & Timmermans, H. J. (2016), 'Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using gps data', *Transportation Planning and Technology* pp. 1–15.

Foell, S., Phithakkitnukoon, S., Kortuem, G., Veloso, M. & Bento, C. (2014), Catch me if you can: Predicting mobility patterns of public transport users, *in* 'Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on', IEEE, pp. 1995–2002.

Forrest, T. L. & Pearson, D. F. (2005), 'Comparison of trip determination methods in household travel surveys enhanced by a global positioning system', *Transportation Research Record* **1917**(1), 63–71.

Fotheringham, A. S. & O'Kelly, M. E. (1989), *Spatial interaction models: formulations and applications*, Vol. 1, Kluwer Academic Publishers Dordrecht.

Fransen, K. & Farber, S. (2019), Using person-based accessibility measures to assess the equity of transport systems, *in* 'Measuring transport equity', Elsevier, pp. 57–72.

Fransen, K., Neutens, T., Farber, S., De Maeyer, P., Deruyter, G. & Witlox, F. (2015), 'Identifying public transport gaps using time-dependent accessibility levels', *Journal of Transport Geography* **48**, 176–187.

Freedom of Information (2015), 'Lu lo dlr interchange values'. Online; Accessed 01/01/2018.
URL: *https://www.whatdotheyknow.com/request/interchange_time_at_london_under*

Freemark, Y. Y. S. (2013), Assessing Journey Time Impacts of Disruptions on London's Piccadilly Line, PhD thesis, Massachusetts Institute of Technology.

Gao, S., Janowicz, K. & Couclelis, H. (2017), 'Extracting urban functional regions from points of interest and human activities on location-based social networks', *Transactions in GIS* **21**(3), 446–467.

Gelfand, A. E., Diggle, P., Guttorp, P. & Fuentes, M. (2010), *Handbook of spatial statistics*, CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Gelman, A. & Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical science* pp. 457–472.

Geurs, K., Kevin, J. & Reggiani, A. (2012), 'Accessibility analysis and transport planning: an introduction', *Accessibility Analysis and Transport Planning. Challenges for Europe and North America. Cheltenham, UK y Northampton, USA: Nectar* pp. 1–14.

Geurs, K. T. (2018), Transport planning with accessibility indices in the netherlands, International Transport Forum Discussion Paper.

Geurs, K. T., Patuelli, R. & Dentinho, T. P. (2016), *Accessibility, Equity and Efficiency: Challenges for Transport and Public Services*, Edward Elgar Publishing.

Geurs, K. T. & Ritsema van Eck, J. (2001), 'Accessibility measures: review and applications. evaluation of accessibility impacts of land-use transportation scenarios, and related social and economic impact', *RIVM rapport 408505006* .

Geurs, K. T. & Van Wee, B. (2004), 'Accessibility evaluation of land-use and transport strategies: review and research directions', *Journal of Transport geography* **12**(2), 127–140.

Geurs, K., Zondag, B., De Jong, G. & de Bok, M. (2010), 'Accessibility appraisal of land-use/transport policy strategies: More than just adding up travel-time savings', *Transportation Research Part D: Transport and Environment* **15**(7), 382–393.

Geweke, J. et al. (1991), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Vol. 196, Federal Reserve Bank of Minneapolis, Research Department.

Gibbs, J. W. (2014), *Elementary principles in statistical mechanics*, Courier Corporation.

Golob, T. F. (2000), 'A simultaneous model of household activity participation and trip chain generation', *Transportation Research Part B: Methodological* **34**(5), 355–376.

Golob, T. F. (2003), 'Structural equation modeling for travel behavior research', *Transportation Research Part B: Methodological* **37**(1), 1–25.

Golob, T. F. & McNally, M. G. (1997), 'A model of activity participation and travel interactions between household heads', *Transportation Research Part B: Methodological* **31**(3), 177–194.

Goodman, A., Jones, A., Roberts, H., Steinbach, R. & Green, J. (2014), "we can all just get on a bus and go': Rethinking independent mobility in the context of the universal provision of free bus travel to young londoners', *Mobilities* **9**(2), 275–293.

Gordon, J. B. (2012), Intermodal passenger flows on London's public transport network: automated inference of full passenger journeys using fare-transaction and vehicle-location data, PhD thesis, Massachusetts Institute of Technology.

Goulias, K. G. & Kitamura, R. (1992), 'Travel demand forecasting with dynamic microsimulation', *Transportation Research Records* **1357**.

Grengs, J. (2015), 'Advancing social equity analysis in transportation with the concept of accessibility'.

Griffin, T. & Huang, Y. (2005), A decision tree classification model to automate trip purpose derivation, *in* 'Proceedings of the ISCA 18th international conference on computer applications in industry and engineering, Sheraton Moana Surfrider, Honolulu, HI', Citeseer, pp. 44–49.

Guagliardo, M. F. (2004), 'Spatial accessibility of primary care: concepts, methods and challenges', *International journal of health geographics* **3**(1), 3.

Haario, H., Saksman, E., Tamminen, J. et al. (2001), 'An adaptive metropolis algorithm', *Bernoulli* **7**(2), 223–242.

Hägerstraand, T. (1970), 'What about people in regional science?', *Papers in regional science* **24**(1), 7–24.

Hägerstrand, T. (1973), 'The domain of human geography', *Directions in geography* pp. 67–87.

Han, G. & Sohn, K. (2016), 'Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model', *Transportation Research Part B: Methodological* **83**, 121–135.

Hananel, R. & Berechman, J. (2016), 'Justice and transportation decision-making: The capabilities approach', *Transport Policy* **49**, 78–85.

Handy, S. L. & Niemeier, D. A. (1997), 'Measuring accessibility: an exploration of issues and alternatives', *Environment and planning A* **29**(7), 1175–1194.

Hansen, W. G. (1959), 'How accessibility shapes land use', *Journal of the American Institute of planners* **25**(2), 73–76.

Hasan, S. & Ukkusuri, S. V. (2014), 'Urban activity pattern classification using topic models from online geo-location data', *Transportation Research Part C: Emerging Technologies* **44**, 363–381.

Hasan, S. & Ukkusuri, S. V. (2017), 'Reconstructing activity location sequences from incomplete check-in data: A semi-markov continuous-time bayesian network model', *IEEE Transactions on Intelligent Transportation Systems* .

Hickman, R., Cao, M., Mella Lira, B., Fillone, A. & Bienvenido Biona, J. (2017), 'Understanding capabilities, functionings and travel in high and low income neighbourhoods in manila', *Social Inclusion* **5**(4), 161–174.

Hoef, J. M. V., Hanks, E. M. & Hooten, M. B. (2017), 'On the relationship between conditional (car) and simultaneous (sar) autoregressive models', *arXiv preprint arXiv:1710.07000* .

Holloway, C. & Tyler, N. (2013), 'A micro-level approach to measuring the accessibility of footways for wheelchair users using the capability model', *Transportation planning and technology* **36**(7), 636–649.

Huang, J. & Ling, C. X. (2005), 'Using auc and accuracy in evaluating learning algorithms', *IEEE Transactions on knowledge and Data Engineering* **17**(3), 299–310.

Huang, L., Li, Q. & Yue, Y. (2010), Activity identification from gps trajectories using spatial temporal pois' attractiveness, *in* 'Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks', ACM, pp. 27–30.

Ingram, D. R. (1971), 'The concept of accessibility: a search for an operational form', *Regional studies* **5**(2), 101–107.

Jahangiri, A. & Rakha, H. A. (2015), 'Applying machine learning techniques to transportation mode recognition using mobile phone sensor data', *Intelligent Transportation Systems, IEEE Transactions on* **16**(5), 2406–2417.

Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T. & Timmermans, H. (2006), 'Integrating bayesian networks and decision trees in a sequential rule-based transportation model', *European Journal of operational research* **175**(1), 16–34.

Jaulmes, R., Pineau, J. & Precup, D. (2005), Active learning in partially observable markov decision processes, *in* 'European Conference on Machine Learning', Springer, pp. 601–608.

Jiang, S., Ferreira, J. & Gonzalez, M. C. (2017), 'Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore', *IEEE Transactions on Big Data* **3**(2), 208–219.

Jordan, M. I. et al. (2004), 'Graphical models', *Statistical Science* **19**(1), 140–155.

Kamruzzaman, M., Yigitcanlar, T., Yang, J. & Mohamed, M. A. (2016), 'Measures of transport-related social exclusion: A critical review of the literature', *Sustainability* **8**(7), 696.

Karlaftis, M. G. & Vlahogianni, E. I. (2011), 'Statistical methods versus neural networks in transportation research: Differences, similarities and some insights', *Transportation Research Part C: Emerging Technologies* **19**(3), 387–399.

Karner, A. (2018), 'Assessing public transit service equity using route-level accessibility measures and public data', *Journal of Transport Geography* **67**, 24–32.

Karner, A. & Niemeier, D. (2013), 'Civil rights guidance and equity analysis methods for regional transportation plans: a critical review of literature and practice', *Journal of Transport Geography* **33**, 126–134.

Karst, T. & van Eck, J. R. R. (2003), 'Evaluation of accessibility impacts of land-use scenarios: the implications of job competition, land-use, and infrastructure developments for the netherlands', *Environment and Planning B: Planning and Design* **30**(1), 69–87.

Keane, M. (1997), 'Current issues in discrete choice modeling', *Marketing Letters* **8**(3), 307–322.

Keleher, L. (2014), 'Sen and nussbaum: Agency and capability-expansion'.

Kelobonye, K., Zhou, H., McCarney, G. & Xia, J. C. (2020), 'Measuring the accessibility and spatial equity of urban services under competition using the cumulative opportunities measure', *Journal of Transport Geography* **85**, 102706.

Kenyon, S. (2003), Understanding social exclusion and social inclusion, *in* 'Proceedings of the Institution of Civil Engineers-Municipal Engineer', Vol. 156, Thomas Telford Ltd, pp. 97–104.

Khisty, C. J. (1996), 'Operationalizing concepts of equity for public project investments', *Transportation Research Record* **1559**(1), 94–99.

Kim, Y., Pereira, F. C., Zhao, F., Ghorpade, A., Zegras, P. C. & Ben-Akiva, M. (2014), Activity recognition for a smartphone based travel survey based on cross-user history data, *in* 'Pattern Recognition (ICPR), 2014 22nd International Conference on', IEEE, pp. 432–437.

Kitamura, R., Chen, C. & Pendyala, R. (1997), 'Generation of synthetic daily activity-travel patterns', *Transportation Research Record: Journal of the Transportation Research Board* (1607), 154–162.

Kobayashi, T., Miller, H. J. & Othman, W. (2011), 'Analytical methods for error propagation in planar space–time prisms', *Journal of Geographical Systems* **13**(4), 327–354.

Koenig, J.-G. (1980), 'Indicators of urban accessibility: theory and application', *Transportation* **9**(2), 145–172.

Koller, D. & Friedman, N. (2009), *Probabilistic graphical models: principles and techniques*, MIT press.

Koushik, A. N., Manoj, M. & Nezamuddin, N. (2020), 'Machine learning applications in activity-travel behaviour research: a review', *Transport reviews* **40**(3), 288–311.

Krawczyk, B. (2016), 'Learning from imbalanced data: open challenges and future directions', *Progress in Artificial Intelligence* **5**(4), 221–232.

Krishnakumar, J. (2007), 'Going beyond functionings to capabilities: An econometric model to explain and estimate capabilities', *Journal of Human Development* **8**(1), 39–63.

Krishnakumar, J. (2013), 'Quantitative methods for the capability approach', *Social and cultural development of human resources* pp. 1–28.

Krishnakumar, J. & Ballon, P. (2008), 'Estimating basic capabilities: A structural equation model applied to bolivia', *World Development* **36**(6), 992–1010.

Kruschke, J. (2010), *Doing Bayesian data analysis: A tutorial introduction with R*, Academic Press.

Kuijpers, B., Miller, H. J., Neutens, T. & Othman, W. (2010), 'Anchor uncertainty and space-time prisms on road networks', *International Journal of Geographical Information Science* **24**(8), 1223–1248.

Kwan, M.-P. (1998), 'Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework', *Geographical analysis* **30**(3), 191–216.

Kwan, M.-P. (1999), 'Gender and individual access to urban opportunities: a study using space–time measures', *The Professional Geographer* **51**(2), 210–227.

Kwan, M.-P. (2013), 'Beyond space (as we knew it): Toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and giscience', *Annals of the Association of American Geographers* **103**(5), 1078–1086.

Kwan, M.-P. & Kotsev, A. (2015), 'Gender differences in commute time and accessibility in s ofia, b ulgaria: a study using 3 d geovisualisation', *The Geographical Journal* **181**(1), 83–96.

Kwan, M.-P., Murray, A. T., O'Kelly, M. E. & Tiefelsdorf, M. (2003), 'Recent advances in accessibility research: Representation, methodology and applications', *Journal of Geographical Systems* **5**(1), 129–138.

Langley, R. B. et al. (1999), 'Dilution of precision', *GPS world* **10**(5), 52–59.

Lawson, A. (2009), *Bayesian disease mapping*, 1 edn, CRC Press.

Lee, J. & Miller, H. J. (2018), 'Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in columbus, ohio, usa', *Applied geography* **93**, 47–63.

Lee, J. & Miller, H. J. (2019), 'Analyzing collective accessibility using average space-time prisms', *Transportation Research Part D: Transport and Environment* **69**, 250–264.

Lee, S. G. & Hickman, M. (2014), 'Trip purpose inference using automated fare collection data', *Public Transport* **6**(1-2), 1–20.

Lelli, S. (2008), 'Operationalising sen's capability approach: The influence of the selected technique', *The Capability approach: Concepts, Measures and Application* pp. 310–361.

Levinson, D. & Chen, W. (2005), Paving new ground: a markov chain model of the change in transportation networks and land use, *in* 'Access to destinations', Emerald Group Publishing Limited, pp. 243–266.

Levinson, D. M. & Kumar, A. (1994), 'The rational locator: why travel times have remained stable', *Journal of the american planning association* **60**(3), 319–332.

Levitas, R., Pantazis, C., Fahmy, E., Gordon, D., Lloyd, E. & Patsios, D. (2007), 'The multi-dimensional analysis of social exclusion'.

Li, R. & Tong, D. (2016), 'Constructing human activity spaces: A new approach incorporating complex urban activity-travel', *Journal of Transport Geography* **56**, 23–35.

Li, T., Guan, H., Ma, J., Zhang, G. & Liang, K. (2017), 'Modeling travel mode choice behavior with bounded rationality using markov logic networks', *Transportation Letters* pp. 1–8.

Liao, F., Rasouli, S. & Timmermans, H. (2014), 'Incorporating activity-travel time uncertainty and stochastic space–time prisms in multistate supernetworks for activity-travel scheduling', *International Journal of Geographical Information Science* **28**(5), 928–945.

Liao, L., Fox, D. & Kautz, H. (2006), Location-based activity recognition, *in* 'Advances in Neural Information Processing Systems', pp. 787–794.

Liao, L., Fox, D. & Kautz, H. (2007), 'Extracting places and activities from gps traces using hierarchical conditional random fields', *The International Journal of Robotics Research* **26**(1), 119–134.

Liao, L., Patterson, D. J., Fox, D. & Kautz, H. (2007), 'Learning and inferring transportation routines', *Artificial Intelligence* **171**(5), 311–331.

Lin, M. & Hsu, W.-J. (2014), 'Mining gps data for mobility patterns: A survey', *Pervasive and Mobile Computing* **12**, 1–16.

Lira, B. M. (2019), Using a capability approach-based survey for reducing equity gaps in transport appraisal: Application in santiago de chile, *in* 'Measuring Transport Equity', Elsevier, pp. 247–264.

Löchl, M. & Axhausen, K. W. (2010), 'Modeling hedonic residential rents for land use and transport simulation while considering spatial effects', *Journal of Transport and Land Use* **3**(2), 39–63.

Long, Y. & Shen, Z. (2015), Discovering functional zones using bus smart card data and points of interest in beijing, *in* 'Geospatial analysis to support urban planning in Beijing', Springer, pp. 193–217.

Long, Y. & Thill, J.-C. (2015), 'Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing', *Computers, Environment and Urban Systems* **53**, 19–35.

López, E., Gutiérrez, J. & Gómez, G. (2008), 'Measuring regional cohesion effects of large-scale transport infrastructure investments: an accessibility approach', *European Planning Studies* **16**(2), 277–301.

Lotfi, S. & Koohsari, M. J. (2009), 'Measuring objective accessibility to neighborhood facilities in the city (a case study: Zone 6 in tehran, iran)', *Cities* **26**(3), 133–140.

Lowe, J. M. & Sen, A. (1996), 'Gravity model applications in health planning: Analysis of an urban hospital market', *Journal of Regional Science* **36**(3), 437–461.

Lucas, C. G. & Kemp, C. (2015), 'An improved probabilistic account of counterfactual reasoning.', *Psychological review* **122**(4), 700.

Lucas, K. (2012), 'Transport and social exclusion: Where are we now?', *Transport policy* **20**, 105–113.

Lucas, K. & Porter, G. (2016), 'Mobilities and livelihoods in urban development contexts: introduction.', *Journal of transport geography.* **55**, 129–131.

Lucas, K., Van Wee, B. & Maat, K. (2016), 'A method to evaluate equitable accessibility: combining ethical theories and accessibility-based approaches', *Transportation* **43**(3), 473–490.

Ma, T.-Y. (2015), 'Bayesian networks for multimodal mode choice behavior modelling: a case study for the cross border workers of luxembourg', *Transportation research procedia* **10**, 870–880.

Ma, T.-Y. & Klein, S. (2018), 'Bayesian networks for constrained location choice modeling using structural restrictions and model averaging', *European Journal of Transport and Infrastructure Research* **18**(1).

Maciel, V., Kuwahara, M., Fronzaglia, M., Scarano, P. & Muramatsu, R. (2015), Accessibility and well-being: Measuring urban (im)mobility as deprivation, *in* '2015 Human Development Capabilities Association Conference'.

Manaugh, K., Badami, M. G. & El-Geneidy, A. M. (2015), 'Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in north america', *Transport policy* **37**, 167–176.

Manley, E. (2016), 'Estimating the topological structure of driver spatial knowledge', *Applied Spatial Analysis and Policy* **9**(2), 165–189.

Martens, K. (2015), 'Accessibility and potential mobility as a guide for policy action', *Transportation research record* **2499**(1), 18–24.

Martens, K., Bastiaanssen, J. & Lucas, K. (2019), Measuring transport equity: Key components, framings and metrics, *in* 'Measuring transport equity', Elsevier, pp. 13–36.

Martens, K. & Golub, A. (2011), Accessibility measures from an equity perspective, *in* 'Bijdrage aan het Colloquium Vervoersplanologisch Speruwerk', Vol. 24.

Martens, K. & Golub, A. (2012), '11. a justice-theoretic exploration of accessibility measures', *Accessibility analysis and transport planning: Challenges for Europe and North America* **195**.

Martínez, F. J. & Araya, C. A. (2000), 'A note on trip benefits in spatial interaction models', *Journal of Regional Science* **40**(4), 789–796.

McFadden, D. (1998), 'Measuring willingness-to-pay for transportation improvements', *Theoretical Foundations of Travel Choice Modeling* **339**, 364.

McGowen, P. & McNally, M. (2007), Evaluating the potential to predict activity types from gps and gis data, *in* 'Transportation Research Board 86th Annual Meeting, Washington', Citeseer.

McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E. & Dibben, C. (2011), 'The english indices of deprivation 2010', *London: Department for Communities and Local Government* .

Merlin, L. A. & Hu, L. (2017), 'Does competition matter in measures of job accessibility? explaining employment in los angeles', *Journal of Transport Geography* **64**, 77–88.

Miller, H. J. (1991), 'Modelling accessibility using space-time prism concepts within geographical information systems', *International Journal of Geographical Information System* **5**(3), 287–301.

Miller, H. J. (2005), Place-based versus people-based accessibility, *in* 'Access to destinations', Emerald Group Publishing Limited, pp. 63–89.

Miller, H. J. (2016), 'Time geography and space–time prism', *International encyclopedia of geography: People, the earth, environment and technology* pp. 1–19.

Mitra, S. (2006), 'The capability approach and disability', *Journal of disability policy studies* **16**(4), 236–247.

Mohamed, K., Côme, E., Baro, J. & Oukhellou, L. (2014), 'Understanding passenger patterns in public transit through smart card and socioeconomic data', *UrbComp,(Seattle, WA, USA)* .

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N. & Axhausen, K. W. (2015), 'Comparison of travel diaries generated from smartphone data and dedicated gps devices', *Transportation Research Procedia* **11**, 227–241.

Moya-Gómez, B., Salas-Olmedo, M. H., García-Palomares, J. C. & Gutiérrez, J. (2018), 'Dynamic accessibility using big data: the role of the changing conditions of network congestion and destination attractiveness', *Networks and Spatial Economics* **18**(2), 273–290.

Mullen, C., Tight, M., Whiteing, A. & Jopson, A. (2014), 'Knowing their place on the roads: What would equality mean for walking and cycling?', *Transportation research part A: policy and practice* **61**, 238–248.

Murphy, K. P. (2012), Undirected graphical models (markov random fields), *in* 'Machine Learning: A Probabilistic Perspective', MIT Press LTD, pp. 661–705.

Nahmias-Biran, B.-h., Martens, K. & Shiftan, Y. (2017), 'Integrating equity in transportation project assessment: A philosophical exploration and its practical implications', *Transport reviews* **37**(2), 192–210.

Nahmias-Biran, B.-h. & Shiftan, Y. (2019), 'Using activity-based models and the capability approach to evaluate equity considerations in transportation projects', *Transportation* pp. 1–19.

National Human Genome Research Institute (2017), 'An overview of the human genome project'. Online; Accessed 24/09/2017.
URL: *https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/*

Neal, R. M. (2000), 'Markov chain sampling methods for dirichlet process mixture models', *Journal of computational and graphical statistics* **9**(2), 249–265.

Neuburger, H. (1971), 'User benefit in the evaluation of transport and land use plans', *Journal of Transport Economics and Policy* pp. 52–75.

Neutens, T. (2015), 'Accessibility, equity and health care: review and research directions for transport geographers', *Journal of Transport Geography* **43**, 14–27.

Neutens, T., Delafontaine, M., Schwanen, T. & Van de Weghe, N. (2012), 'The relationship between opening hours and accessibility of public service delivery', *Journal of Transport Geography* **25**, 128–140.

Neutens, T., Delafontaine, M., Scott, D. M. & De Maeyer, P. (2012), 'An analysis of day-to-day variations in individual space–time accessibility', *Journal of Transport Geography* **23**, 81–91.

Neutens, T., Schwanen, T. & Witlox, F. (2011), 'The prism of everyday life: Towards a new research agenda for time geography', *Transport reviews* **31**(1), 25–47.

Neutens, T., Schwanen, T., Witlox, F. & De Maeyer, P. (2010), 'Equity of urban service delivery: a comparison of different accessibility measures', *Environment and Planning A* **42**(7), 1613–1635.

Neutens, T., Versichele, M. & Schwanen, T. (2010), 'Arranging place and time: A gis toolkit to assess person-based accessibility of urban opportunities', *Applied Geography* **30**(4), 561–575.

Noble, M., McLennan, D., Wilkinson, K., Whitworth, A., Exley, S., Barnes, H., Dibben, C., McLennan, D. et al. (2007), 'The english indices of deprivation 2007'.

Nordbakke, S. (2013), 'Capabilities for mobility among urban older women: barriers, strategies and options', *Journal of transport geography* **26**, 166–174.

Noulas, A., Scellato, S., Mascolo, C. & Pontil, M. (2011), 'An empirical study of geographic user activity patterns in foursquare.', *ICwSM* **11**, 70–573.

Nutley, S. (1998), 'Rural areas: the accessibility problem', *Modern Transport Geography, 2nd rev. ed., Wiley and sons, Chichester* pp. 185–215.

Omrani, H. (2015), 'Predicting travel mode of individuals by machine learning', *Transportation Research Procedia* **10**, 840–849.

Ordnance Survey (2012), 'Points of interest classification scheme'.

Ordnance Survey (2018), 'Points of interest user guide and technical specification'.

Orr, S. (2010), Evaluation of transport accessibility for elderly and disabled people: a proposal for an activity-based quality of life approach, *in* 'European Transport Conference, 2010Association for European Transport'.

Ortega-Tong, M. A. (2013), Classification of London's public transport users using smart card data, PhD thesis, Massachusetts Institute of Technology.

Páez, A., Gertes Mercado, R., Farber, S., Morency, C. & Roorda, M. (2010), 'Relative accessibility deprivation indicators for urban settings: definitions and application to food deserts in montreal', *Urban Studies* **47**(7), 1415–1438.

Páez, A., Scott, D. M. & Morency, C. (2012), 'Measuring accessibility: positive and normative implementations of various accessibility indicators', *Journal of Transport Geography* **25**, 141–153.

Papa, E., Coppola, P., Angiello, G. & Carpentieri, G. (2017), 'The learning process of accessibility instrument developers: Testing the tools in planning practice', *Transportation Research Part A: Policy and Practice* **104**, 108–120.

Pardo, L. (2005), *Statistical inference based on divergence measures*, CRC Press.

Patterson, D. J., Liao, L., Fox, D. & Kautz, H. (2003), Inferring high-level behavior from low-level sensors, *in* 'UbiComp 2003: Ubiquitous Computing', Springer, pp. 73–89.

Patterson, Z. & Farber, S. (2015), 'Potential path areas and activity spaces in application: a review', *Transport Reviews* **35**(6), 679–700.

Paulssen, M., Temme, D., Vij, A. & Walker, J. L. (2014), 'Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice', *Transportation* **41**(4), 873–888.

Pearl, J. (2009), *Introduction to Probabilities, Graphs, and Causal Models*, Cambridge University Press, p. 1–40.

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**(4), 557–568.

Pereira, R. H. (2019), 'Future accessibility impacts of transport policy scenarios: Equity and sensitivity to travel time thresholds for bus rapid transit expansion in rio de janeiro', *Journal of Transport Geography* **74**, 321–332.

Pereira, R. H., Schwanen, T. & Banister, D. (2017), 'Distributive justice and equity in transportation', *Transport Reviews* **37**(2), 170–191.

Perrakis, K., Karlis, D., Cools, M. & Janssens, D. (2015), 'Bayesian inference for transportation origin–destination matrices: the poisson–inverse gaussian and other poisson mixtures', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(1), 271–296.

Perrakis, K., Karlis, D., Cools, M., Janssens, D., Vanhoof, K. & Wets, G. (2012), 'A bayesian approach for modeling origin–destination matrices', *Transportation Research Part A: Policy and Practice* **46**(1), 200–212.

Pirie, G. (1981), 'The possibility and potential of public policy on accessibility', *Transportation Research Part A: General* **15**(5), 377–381.

Popkowski Leszczyc, P. T. & Timmermans, H. J. (2002), 'Unconditional and conditional competing risk models of activity duration and activity sequencing decisions: An empirical comparison', *Journal of Geographical systems* **4**(2), 157–170.

Pred, A. (1977), 'The choreography of existence: comments on hägerstrand's time-geography and its usefulness', *Economic geography* pp. 207–221.

Preston, J. & Rajé, F. (2007), 'Accessibility, mobility and transport-related social exclusion', *Journal of Transport Geography* **15**(3), 151–160.

Pyrialakou, V. D., Gkritza, K. & Fricker, J. D. (2016), 'Accessibility, mobility, and realized travel behavior: Assessing transport disadvantage from a policy perspective', *Journal of transport geography* **51**, 252–269.

Ramjerdi, F. (2006), 'Equity measures and their performance in transportation', *Transportation Research Record* **1983**(1), 67–74.

Rashid, K., Yigitcanlar, T. & Bunker, J. M. (2010), 'Minimising transport disadvantage to support knowledge city formation: applying the capability approach to select indicators', *Melbourne 2010 Knowledge Cities World Summit: 3rd Knowledge Cities World Summit.* .

Rasouli, S. & Timmermans, H. J. (2014), 'Uncertain travel times and activity schedules under conditions of space-time constraints and invariant choice heuristics', *Environment and Planning B: Planning and Design* **41**(6), 1022–1030.

Rawls, J. (2009), *A theory of justice*, Harvard university press.

Reades, J. (2014), 'Tfl data source documentation', Personal communication.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M. & Srivastava, M. (2010), 'Using mobile phones to determine transportation modes', *ACM Transactions on Sensor Networks (TOSN)* **6**(2), 13.

Robert, C. (2014), 'Machine learning, a probabilistic perspective'.

Roberto, E. (2015), 'Measuring inequality and segregation', *arXiv preprint arXiv:1508.01167* pp. 1–26.

Robeyns, I. (2005*a*), 'The capability approach: a theoretical survey', *Journal of human development* **6**(1), 93–117.

Robeyns, I. (2005*b*), 'Selecting capabilities for quality of life measurement', *Social indicators research* **74**(1), 191–215.

Rue, H. & Held, L. (2005), *Gaussian Markov random fields: theory and applications*, CRC Press.

Ryan, J., Wretstrand, A. & Schmidt, S. M. (2015), 'Exploring public transport as an element of older persons' mobility: A capability approach perspective', *Journal of transport geography* **48**, 105–114.

Santos, B., Antunes, A. & Miller, E. J. (2008), 'Integrating equity objectives in a road network design model', *Transportation Research Record* **2089**(1), 35–42.

Sari Aslam, N., Cheng, T. & Cheshire, J. (2019), 'A high-precision heuristic model to detect home and work locations from smart card data', *Geo-spatial Information Science* **22**(1), 1–11.

Sarkar, C., Webster, C., Pryor, M., Tang, D., Melbourne, S., Zhang, X. & Jianzheng, L. (2015), 'Exploring associations between urban green, street design and walking: Results from the greater london boroughs', *Landscape and Urban Planning* **143**, 112–125.

Schmöcker, J.-D., Quddus, M. A., Noland, R. B. & Bell, M. G. (2008), 'Mode choice of older and disabled people: a case study of shopping trips in london', *Journal of Transport Geography* **16**(4), 257–267.

Schönfelder, S. (2001), Some notes on space, location and travel behaviour, *in* 'Swiss Transport Research Conference'.

Schönfelder, S. & Axhausen, K. W. (2003), 'Activity spaces: measures of social exclusion?', *Transport policy* **10**(4), 273–286.

Sclar, E. D., Lönnroth, M. & Wolmar, C. (2014), *Urban access for the 21st century: Finance and governance models for transport infrastructure*, Routledge.

Sen, A. (1992), *Inequality reexamined*, Oxford University Press.

Sen, A. (2008), 'The idea of justice', *Journal of human development* **9**(3), 331–342.

Sen, A. (2014), 'Development as freedom (1999)', *Roberts, JT, Hite, AB & Chorev, N. The Globalization and Development Reader: Perspectives on Development and Global Change* **2**, 525–547.

Sen, A. & Smith, T. E. (2012), *Gravity models of spatial interaction behavior*, Springer Science & Business Media.

Sen, A. et al. (1990), 'Development as capability expansion', *Human development and the international development strategy for the 1990s* **1**.

Shafique, M. A. & Hato, E. (2015), 'Use of acceleration data for transportation mode prediction', *Transportation* **42**(1), 163–188.

Shen, L. & Stopher, P. R. (2013), 'A process for trip purpose imputation from global positioning system data', *Transportation Research Part C: Emerging Technologies* **36**, 261–267.

Shen, L. & Stopher, P. R. (2014), 'Review of gps travel survey and gps data-processing methods', *Transport Reviews* **34**(3), 316–334.

Shen, Q. (1998), 'Location characteristics of inner-city neighborhoods and employment accessibility of low-wage workers', *Environment and planning B: Planning and Design* **25**(3), 345–365.

Simma, A. & Axhausen, K. (2003), *Interactions between travel behaviour, accessibility and personal characteristics: The case of the Upper Austria Region*, ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme.

Smardon, R. C. (1988), 'Perception and aesthetics of the urban environment: Review of the role of vegetation', *Landscape and Urban Planning* **15**(1-2), 85–106.

Smith, C., Quercia, D. & Capra, L. (2012), Anti-gravity underground, *in* 'the 2nd Workshop on Pervasive Urban Applications (PURBA)'.

Smith, N., Hirsch, D. & Davis, A. (2012), 'Accessibility and capability: the minimum transport needs and costs of rural households', *Journal of Transport Geography* **21**, 93–101.

Social Exclusion Unit (2003), 'Making the connections: final report on transport and social exclusion', *http://webarchive.nationalarchives.gov. uk* .

Song, X., Zhang, Q., Sekimoto, Y. & Shibasaki, R. (2014), Prediction of human emergency behavior and their mobility following large-scale disaster, *in* 'Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 5–14.

Steinbach, R., Green, J. & Edwards, P. (2012), 'Look who's walking: Social and environmental correlates of children's walking in london', *Health & place* **18**(4), 917–927.

Stenneth, L., Wolfson, O., Yu, P. S. & Xu, B. (2011), Transportation mode detection using mobile phones and gis information, *in* 'Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems', ACM, pp. 54–63.

Stępniak, M. & Goliszek, S. (2017), Spatio-temporal variation of accessibility by public transport—the equity perspective, *in* 'The rise of big spatial data', Springer, pp. 241–261.

Stopher, P. R. & Greaves, S. P. (2007), 'Household travel surveys: Where are we going?', *Transportation Research Part A: Policy and Practice* **41**(5), 367–381.

Sun, F.-T., Yeh, Y.-T., Cheng, H.-T., Kuo, C. & Griss, M. (2014), Nonparametric discovery of human routines from sensor data, *in* 'Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on', IEEE, pp. 11–19.

Swait, J. (2001), 'Choice set generation within the generalized extreme value family of discrete choice models', *Transportation Research Part B: Methodological* **35**(7), 643–666.

TfL (2014), 'Understanding the travel needs of london's diverse communities: A summary of existing research'. Online; Accessed 07/10/2015.
URL: *https://tfl.gov.uk/cdn/static/cms/documents/understanding-the-travel-needs-of-london-diverse-communities.pdf*

Titheridge, H., Achuthan, K., Mackett, R. & Solomon, J. (2009), 'Assessing the extent of transport social exclusion among the elderly'.

Titheridge, H., Mackett, R., Christie, N., Oviedo Hernández, D. & Ye, R. (2014), 'Transport and poverty: a review of the evidence'.

Tomarken, A. J. & Baker, T. B. (2003), 'Introduction to the special section on structural equation modeling.', *Journal of Abnormal Psychology* **112**(4), 523.

Tong, L., Zhou, X. & Miller, H. J. (2015), 'Transportation network design for maximizing space–time accessibility', *Transportation Research Part B: Methodological* **81**, 555–576.

Train, K. E. (2009), *Discrete choice methods with simulation*, Cambridge university press.

Transport for London (2010), 'Measuring public transport accessibility levels ptals summary'. Online; accessed 16/09/2014.
URL: *http://data.london.gov.uk/documents/PTAL-methodology.pdf*

Transport for London (2011), 'Travel in london,supplementary report:london travel demandsurvey (ltds)'. Online; Accessed 29/03/2016.
URL: *http://content.tfl.gov.uk/london-travel-demand-survey.pdf*

Transport for London (2019), 'Public transport journeys by type of transport'. Online; Accessed 06/01/2019.

URL: *https://data.london.gov.uk/dataset/public-transport-journeys-type-transport*

Tyler, N. (2006), 'Capabilities and radicalism: Engineering accessibility in the 21st century', *Transportation planning and technology* **29**(5), 331–358.

UK Government (2015), 'English indices of deprivation 2015'. Online; Accessed 24/09/2016.
URL: *https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015*

UKGovenrment (2018), 'Press release record boost to r and d and new transport fund to help build economy fit for the future'. Online; Accessed 15/1/2018.
URL: *https://www.gov.uk/government/news/record-boost-to-rd-and-new-transport-fund-to-help-build-economy-fit-for-the-future*

Uniman, D., Attanucci, J., Mishalani, R. & Wilson, N. (2010), 'Service reliability measurement using automated fare card data: Application to the london underground', *Transportation Research Record: Journal of the Transportation Research Board* (2143), 92–99.

Van Acker, V., Van Wee, B. & Witlox, F. (2010), 'When transport geography meets social psychology: toward a conceptual model of travel behaviour', *Transport Reviews* **30**(2), 219–240.

Van Acker, V., Witlox, F. & Van Wee, B. (2007), 'The effects of the land use system on travel behavior: a structural equation modeling approach', *Transportation planning and technology* **30**(4), 331–353.

Van Wee, B. (2011), *Transport and ethics: ethics and the evaluation of transport policies and projects*, Edward Elgar Publishing.

Van Wee, B. (2016), 'Accessible accessibility research challenges', *Journal of transport geography* **51**, 9–16.

Van Wee, B. & Geurs, K. (2011), 'Discussing equity and social exclusion in accessibility evaluations', *EJTIR* **11**(4), 350–367.

Van Wee, B. & Roeser, S. (2013), 'Ethical theories and the cost–benefit analysis-based ex ante evaluation of transport policies and plans', *Transport reviews* **33**(6), 743–760.

Verhoeven, M., Arentze, T., Timmermans, H. J. & Van Der Waerden, P. (2005), 'Modeling the impact of key events on long-term transport mode choice decisions: decision network approach using event history data', *Transportation Research Record* **1926**(1), 106–114.

Vovsha, P. (1997), 'Application of cross-nested logit model to mode choice in tel aviv, israel, metropolitan area', *Transportation Research Record* **1607**(1), 6–15.

Wang, C.-H. & Chen, N. (2015), 'A gis-based spatial statistical approach to modeling job accessibility by transportation mode: case study of columbus, ohio', *Journal of transport geography* **45**, 1–11.

Wang, F. (2000), 'Modeling commuting patterns in chicago in a gis environment: A job accessibility perspective', *The Professional Geographer* **52**(1), 120–133.

Wang, W., Attanucci, J. P. & Wilson, N. H. (2011), 'Bus passenger origin-destination estimation and related analyses using automated data collection systems', *Journal of Public Transportation* **14**(4), 7.

Wang, Y., Chen, B. Y., Yuan, H., Wang, D., Lam, W. H. & Li, Q. (2018), 'Measuring temporal variation of location-based accessibility using space-time utility perspective', *Journal of Transport Geography* **73**, 13–24.

Wang, Y., de Almeida Correia, G. H., de Romph, E. & Timmermans, H. (2017), 'Using metro smart card data to model location choice of after-work activities: An application to shanghai', *Journal of Transport Geography* **63**, 40–47.

Wegener, M., Eskelinnen, H., Fürst, F., Schürmann, C. & Spiekermann, K. (2000), 'Indicators of geographical position', *Final Report of the Working Group "Geographical Position" of the Study Programme on European Spatial Planning. Dortmund, IRPUD* .

Weibull, J. W. (1976), 'An axiomatic approach to the measurement of accessibility', *Regional science and urban economics* **6**(4), 357–379.

Wen, C.-H., Wang, W.-C. & Fu, C. (2012), 'Latent class nested logit model for analyzing high-speed rail access mode choice', *Transportation Research Part E: Logistics and Transportation Review* **48**(2), 545–554.

Wermuth, M., Sommer, C. & Kreitz, M. (2003), 'Impact of new technologies in travel surveys', *Transport survey quality and innovation* pp. 455–482.

Widhalm, P., Nitsche, P. & Brändie, N. (2012), Transport mode detection with realistic smartphone sensor data, *in* 'Pattern Recognition (ICPR), 2012 21st International Conference on', IEEE, pp. 573–576.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S. & González, M. C. (2015), 'Discovering urban activity patterns in cell phone data', *Transportation* **42**(4), 597–623.

Wilson, A. G. (1971), 'A family of spatial interaction models, and associated developments', *Environment and Planning A* **3**(1), 1–32.

Winter, S. & Yin, Z.-C. (2011), 'The elements of probabilistic time geography', *GeoInformatica* **15**(3), 417–434.

Witlox, F. (2015), 'Beyond the data smog?', *Transport Reviews* **35**(3), 245–249.

Wixey, S., Jones, P., Lucas, K. & Aldridge, M. (2005), 'Measuring accessibility as experienced by different socially disadvantaged groups', *London, Transit Studies Group, University of Westminster* .

Wolf, J., Loechl, M., Thompson, M. & Arce, C. (2003), 'Trip rate analysis in gps-enhanced personal travel surveys', *Transport survey quality and innovation* **28**, 483–98.

Wolff, J. (2007), 'Equality: The recent history of an idea', *Journal of Moral Philosophy* **4**(1), 125–136.

World Bank Group (2005), 'Introduction to poverty analysis'. Online; Accessed 15/10/2019.
URL: *http://siteresources.worldbank.org/PGLP/Resources/PovertyManual.pdf*

Wright, S. (1921), 'Correlation and causation', *Journal of agricultural research* **20**(7), 557–585.

Wu, B. M. & Hine, J. P. (2003), 'A ptal approach to measuring changes in bus service accessibility', *Transport Policy* **10**(4), 307–320.

Wu, L., Yang, B. & Jing, P. (2016), 'Travel mode detection based on gps raw data collected by smartphones: a systematic review of the existing methodologies', *Information* **7**(4), 67.

Wu, Y.-H. & Miller, H. J. (2001), 'Computational tools for measuring space-time accessibility within dynamic flow transportation networks', *Journal of Transportation and Statistics* **4**(2/3), 1–14.

Xia, N., Cheng, L., Chen, S., Wei, X., Zong, W. & Li, M. (2018), 'Accessibility based on gravity-radiation model and google maps api: A case study in australia', *Journal of Transport Geography* **72**, 178–190.

Xiao, G., Juan, Z. & Zhang, C. (2015), 'Travel mode detection based on gps track data and bayesian networks', *Computers, Environment and Urban Systems* **54**, 14–22.

Xiao, G., Juan, Z. & Zhang, C. (2016), 'Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization', *Transportation Research Part C: Emerging Technologies* **71**, 447–463.

Xie, C. & Waller, S. (2010), 'Estimation and application of a bayesian network model for discrete travel choice analysis', *Transportation Letters* **2**(2), 125–144.

Xiong, C., Chen, X., He, X., Guo, W. & Zhang, L. (2015), 'The analysis of dynamic travel mode choice: a heterogeneous hidden markov approach', *Transportation* **42**(6), 985–1002.

Xiong, C., Hetrakul, P. & Zhang, L. (2014), 'On ride-sharing: a departure time choice analysis with latent carpooling preference', *Journal of Transportation Engineering* **140**(8), 04014033.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z. & Li, Q. (2015), 'Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach', *Transportation* **42**(4), 625–646.

Yamamoto, T., Kitamura, R. & Fujii, J. (2002), 'Drivers' route choice behavior: analysis by data mining algorithms', *Transportation Research Record* **1807**(1), 59–66.

Yamamoto, T., Kitamura, R. & Pendyala, R. M. (2004), 'Comparative analysis of time-space prism vertices for out-of-home activity engagement on working and nonworking days', *Environment and Planning B* **31**(2), 235–250.

Yang, Q. (2009), Activity recognition: linking low-level sensors to high-level intelligence., *in* 'IJCAI', Vol. 9, pp. 20–25.

Yang, X. & Day, J. (2016), Operationalizing the capabilities approach for urban policy evaluation: The travel welfare impacts of government job resettlement, *in* 'Geography Research Forum', Vol. 35, pp. 113–137.

Ye, J., Zhu, Z. & Cheng, H. (2013), What's your next move: User activity prediction in location-based social networks, *in* 'Proceedings of the SIAM International Conference on Data Mining. SIAM', SIAM.

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F. & Pozdnoukhov, A. (2017), 'A generative model of urban activities from cellular data', *IEEE Transactions on Intelligent Transportation Systems* **19**(6), 1682–1696.

Yin, M., Sheehan, M., Feygin, S., Paiement, J.-F. & Pozdnoukhov, A. (2018), 'A generative model of urban activities from cellular data', *IEEE Transactions on Intelligent Transportation Systems* **19**(6), 1682–1696.

Yuan, J., Zheng, Y. & Xie, X. (2012), Discovering regions of different functions in a city using human mobility and pois, *in* 'Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 186–194.

Yun, D.-S., Lee, J.-Y. & Sinha, K. C. (2000), 'Modeling prework trip-making and home departure time choice', *Journal of transportation engineering* **126**(4), 308–312.

Zhang, F., Jin, B., Ge, T., Ji, Q. & Cui, Y. (2016), Who are my familiar strangers?: Revealing hidden friend relations and common interests from smart card data, *in* 'Proceedings of the 25th ACM International on Conference on Information and Knowledge Management', ACM, pp. 619–628.

Zhang, L., Liu, L., Bao, S., Qiang, M. & Zou, X. (2015), 'Transportation mode detection based on permutation entropy and extreme learning machine', *Mathematical Problems in Engineering* **2015**.

Zhao, J., Rahbee, A. & Wilson, N. H. (2007), 'Estimating a rail passenger trip origin-destination matrix using automatic data collection systems', *Computer-Aided Civil and Infrastructure Engineering* **22**(5), 376–387.

Zheng, Y. (2015), 'Trajectory data mining: an overview', *ACM Transactions on Intelligent Systems and Technology (TIST)* **6**(3), 29.

Zheng, Y., Liu, L., Wang, L. & Xie, X. (2008), Learning transportation mode from raw gps data for geographic applications on the web, *in* 'Proceedings of the 17th international conference on World Wide Web', ACM, pp. 247–256.

Zhu, Z., Chen, X., Xiong, C. & Zhang, L. (2018), 'A mixed bayesian network for two-dimensional decision modeling of departure time and mode choice', *Transportation* **45**(5), 1499–1522.

Zucchini, W. & MacDonald, I. L. (2009), *Hidden Markov models for time series: an introduction using R*, CRC Press.