A Framework for Big Data in Urban Mobility and Movement Patterns Analysis

Eusebio Amechi Odiari

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy (PhD)

The University of Leeds
Faculty of Environment, School of Geography

September, 2018

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to thank my supervisors Prof Mark Birkin, Prof Susan Grant-Muller and Dr Nicolas Malleson for the fervent guidance they have given me throughout the research work. This work was supported by a grant from the Consumer Data Research Centre (CDRC), Economic and Social Research Council (ESRC) and by funding from the Association of Train Operating Companies (ATOC). I am indebted to Tony Magee and Nick Wilson of ATOC for scoping the project in, for their encouragement and technical support.

This project has been a multi-disciplinary one, cutting across schools within the Faculty of Environment (School of Geography and Institute for Transport Studies), the Leeds Institute for Data Analytics (LIDA) and the project industrial sponsors the Association of Train Operating Companies (ATOC). Managing the project would not have been possible without the unique PhD administrative structure within the School of Geography, University of Leeds, which includes the project support team Dr Nik Lomax and the independent assessor Dr Gordon Mitchell. I am particularly grateful for their technical support, research advice and industrial guidance.

I would like to acknowledge the moral support, the several technically inciting discussions, and the generosity I have enjoyed from colleagues at the Consumer Data Research Centre (CDRC), Leeds Institute for Data Analytics. They include in no particular order Emily Sheard, Michelle Morris, Rachel Oldroyd, Robin Lovelace, Nick Hood, Stephen Clark, Tom Waddington, and Phani Kumar. The same acknowledgement goes to postgraduate colleagues at School of Geography, Institute for Transport Studies (ITS), and to teaching, research and support staff at School of Geography and the CDRC.

Abstract

Novel large consumer datasets (called ‘Big Data’) are increasingly readily available. These datasets are typically created for a particular purpose, and as such are skewed, and further do not have the broad spectrum of attributes required for their wider application. Railway ticket data are an example of consumer data, which often have little or no supplementary information about the passengers who purchase them, or the context in which the ticket was used (like crowding-level in the train). These gaps in consumer data present challenges in using these data for planning, and inference on the drivers of mobility choice.

Heckman’s in-depth discussion of ‘sample selection’ bias and ‘omitted variables’ bias (Heckman, 1977), and Rubin’s seminal paper on ‘missing values’ (Rubin, 1976) laid the framework for addressing omitted variables and missing data problems today. On the strength of these, a powerful set of complementary concerted methodologies are developed to harness railways consumer (ticketing) data. A novel spatial microsimulation methodology suitable for skewed interaction data was developed to combine LENNON ticketing, National Rail Travel Survey, and Census interaction data, to yield an attribute-rich micro-population. The micro-population was used as input to a GIS network, logistically constrained by the transit feed specification (GTFS). This identifies the context of passenger mobility. Bayesian models then enable the identification of passenger behaviour, like missing daily trip rates with season tickets, and flows to group stations.

Case studies using the micro-level synthetic data reveal a mechanism of rail-heading phenomena in West Yorkshire, and the impact of a new station at Kirkstall Forge. The spatial microsimulation and GIS-GTFS methods are potentially useful to network operators for the management and maintenance on the railways. The representativeness of the micro-level population created has the potential to alter multi-agent transport simulation genres, by precluding the need for the complexities of utility-maximizing traffic assignment.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	**Brief description**	**Page of first occurrence**
1D	1-dimensional	82
2D	2-dimensional	82
3-D (3D)	3-dimensional	82
ABM	Agent based models	36
ANPR	Automatic Number Plate Recognition	269
ARC	Advanced Research Computing	220
ArcGIS	Environmental systems Research Institute (ESRI) GIS Software	154
ATOC	Association of Train Operating Companies	2
BB	Blackburn Postcode Area	93
BD	Bradford Postcode Area	93
BI	Bayesian Imputation	181
BIY	Bingley Train station Code	67
BR	British Rail	3
BUGS	Bayesian Inference Using Gibbs Sampling	15
BYM	Besag-York-Mollie model convolution	11
CAR	Conditionally Autoregressive	11
CCTV	Close-Circuit Televisions	3
CDRC	Consumer Data Research Centre	2
CI	Credible Intervals	221
CRG	Cross Gates Train station Code	68
CS	Cross-Section	45
DAG	directed acyclic graphs	193
DCM	Discrete Choice Models	35
D-Code	Destination Station Code (for example BIY, LDS)	67
DfT	Department for Transport UK	155
DN	Doncaster Postcode Area	93
EC	European Commission	269
EEA	European Economic Area	271
EIR	Environmental Information Regulations	269
EM	Expectation Maximisation	182
EMB	Expectation-Maximisation with Bootstrapping	40
EMME	Multimodal Equilibrium -Transportation	35

	planning software	
ESRC	Economic and Social Research Council	271
ESRI	Environmental Systems Research Institute	157
ESS	Effective sample Size	217
GA	Genetic Algorithm Spatial Microsimulation	81
GDPR	General Data Protection Regulation	247
GIS	Geographical Information System	8
GPS	Global Positioning System	3
GS	Gibbs sampler based population synthesis (used interchangeably with Gibbs Sampling)	115
GS	Gibbs sampling (used interchangeably with MCMC Gibbs sampler based population synthesis)	115
GTFS	General Transit Feed Specification	11
GWR	Geographically Weighted Regression	240
HC	Hill Climbing	81
HD	Huddersfield Postcode Area	93
HDI	High Density Interval	219
HG	Harrogate Postcode Area	93
HX	Halifax Postcode Area	93
ICT	Information and Communication technology	3
ID	Individual Identity	52
IDW	Inverse Distance Weighting	177
iid	Independent identically distributed	45
IMRAD	Introduction, Method, Results, and Discussion format for scientific articles	16
INSPIRE	Infrastructure for Spatial Information in Europe	269
IPF	Iterative Proportional Fitting	13
ITA	Integrated Transport Authority	45
ITS	Institute for Transport Studies	ii
KEI	Keighley Train station Code	68
KNN	K-nearest neighbour	179
LDS	Leeds Train station Code	67
LENNON	Latest Earnings Networked Nationally Over Night	4
LIDA	Leeds Institute for Data Analytics	ii
L-R	From Left to Right	102

LS	Leeds Postcode area	93
LSOA	Lower Layer Super Output Areas	58
MAR	Missing at random	12
MATSim	Multi-Agent Transport Simulation Toolkit	35
MAUP	Modifiable Areal Unit Problem	61
MC	Monte Carlo	15
M-C	Markov chains	221
MCAR	Missing Completely at Random	24
MCMC	Markov chain Monte Carlo	15
MCSE	Monte Carlo Standard Error	219
MI	Multiple Imputation	181
MICE	Multiple Imputation by Chained Equations	40
m-IPF	Multi-dimensional Iterative Proportional Fitting	13
ML	Maximum Likelihood	181
MNAR	Missing Not-At-Random	24
MOIRA	model of inter-regional activity	144
MSOA	Middle Layer Super Output Areas	60
MTA	Train Ticket of 30-60 days validity	52
MTH	Train Ticket of 180-359 days validity	205
MTL	Train Ticket of 90-180 days validity	41
NRTS	National Rail Travel Survey	10
NS-Sec	National Statistics Socio-economic Classification	60
NTC	Non-commute by Train Passengers	92
NTM	National Transport Model – UK	151
OA	Output Areas	58
O-Code	Origin station Code (for example BIY, LDS)	67
OD	Origin-Destination	51
ODM	origin-destination matrix	144
OL	Oldham Postcode Area	93
ONS	Office for National Statistics	29
ORR	Office for Rail Regulation’s (ORR), now called Office of Rail and Road (ORR)	186
OS	Ordnance Survey	157
PDFH	Passenger Demand Forecasting Handbook	32
PTA	Passenger Transport Authority	186
PTE	Passenger Transport Executives	19
R^2	R-Square value	59
RAM	Random Access Memory	147

RF	Random Forest	179
RSP	Rail Settlement Plan	178
S	Sheffield Postcode Area	93
SA	Simulated Annealing	81
SAIE	Small Area Income Estimates of the Office for National Statistics UK	62
SAR	Sample of Anonymized Records	28
S-B	Simulation-based Population Synthesis	81
SCENES	Scenarios for European Transport - Europe	151
SDG	Steer Davies Gleave Group	19
SMILE	Strategic Model for Integrated Logistic Evaluation - Dutch	151
SON	Steeton & Silsden Train station Code	68
SQL	Structured Query Language	157
STAN	Strategic Planning of Freight Transportation	151
STD	Standard Train Ticket	52
SVM	Support Vector Machine	179
TAE	Total Absolute Error	102
TB	Terra Byte	220
TC	Commute by Train Passengers	89
TCI	Transport Coverage Index	168
TEM	Transport Economics Model - Dutch	151
TfGM	Transport for Greater Manchester	159
TfL	Transport for London	186
TOC	Train Operating Company	2
TRANSIMS	Transportation Analysis and Simulation System	37
TS-CS	Time Series Cross Section	45
UK	United Kingdom	2
UKSA	United Kingdom Statistics Authority	269
USA	United States of America	28
WF	Wakefield Postcode Area	82
WKF	Wakefield Westgate Train Station Code	140
WYCA	West Yorkshire Combined Authority	45
YO	York Postcode Area	93

PART I

This part sets the scene of the thesis, introducing the aims, objectives, research questions, hypothesis, and outlines the structure of the thesis. The literature on the concepts behind the work presented in the thesis is reviewed. This part further describes the setting of the study area, the data used in the research, and presents preliminary pre-processing of such data, as is typical in research of this nature.

Chapter 1
INTRODUCTION AND OVERVIEW

This research has taken place within one of the research themes within the Consumer Data Research Centre (CDRC) at University of Leeds, in collaboration with the Association of Train Operating Companies (ATOC). The ideas of the project that align with ATOC's interest are in exploring the opportunities of using ATOC data to develop new analytical methods for understanding, simulating and predicting urban mobility and individual movement patterns of railway passengers. The aim is to explore opportunities of integrating consumer dataset within the railways with other relevant datasets, to construct a more detailed picture of passenger numbers, journeys and to relate these through spatial analysis to other drivers of mobility. Consumer datasets available in the rail sector are ticketing data. These however, are measured for a particular purpose, to accrue sales revenue for appropriation to train operating companies (TOCs). In the UK, there are about 5m daily ticket sales, representing a large consumer dataset. The ticketing data are also comprehensive with coverage of the entire UK, with about 756 different ticket products representing the rich heterogeneity of passenger ticket use. These qualities of consumer ticketing data result in them being described as 'big data', with the potential to revolutionize our understanding of daily urban phenomena (Kitchin, 2014) if they are harnessed appropriately. As a result of this potential, analysts tend to want to exploit large consumer datasets fully, often beyond what they were originally intended. This extended use often raises a number of issues.

The issues include that, in the case of the rail sector, ticketing data often have little or no supplementary information about the passengers who purchase them, such as socio-demographic attributes, journey purpose, access and egress modes etc. Such information are necessary in order to identify particular drivers of mobility. Furthermore, the ticketing data does not have information about the context in which the ticket was used, such as the level of crowding in the train when the ticket was used, or the waiting times the passenger endured (such information are purchasing factors for passengers who regularly travel on the service). Such information would be necessary for causal inference on passenger mobility behaviour. Other issues with railway ticketing data relate to the daily rate of use of season tickets, the tenure of use of flexible period tickets, and the passengers particular choice of railway station to use, when stations are described as belonging to the same group on the ticket product (for example Wakefield Westgate and Wakefield Kirkgate stations form a group called Wakefield BR[1]). These latter issues relate to idiosyncrasies of passengers behaviour, and a lack of understanding of this behaviour presents challenges in using these data for planning and management of the railways.

In addition, consumer ticketing data which were originally generated for the particular purpose reveal information on a subset of the wider population who consume a particular digital service. This data are as such skewed (and further do not have the broad spectrum of attribute values required for their wider application). To illustrate the skew, for example, in the rail sector, ticketing data consists of ~60% commuters, whilst the wider Census population consists of ~10% commuters, making ticketing data a skewed (or biased) subset of the wider Census population. All the issues with consumer ticketing data are typical of the type of gaps in novel sources of big data (like retail reward cards, twitter data, and network GPS traces, etc.), and these gaps have to be plugged prior to the use of such data beyond the original intention in measuring such data. In the case of ticketing data originally intended for revenue accumulation, the gaps have to be plugged prior to using such data for detailed mobility analysis.

1.1 Background

Urban mobility and movement patterns has been the subject of studies for over a century (Morris, 2013), as an understanding of the rich tapestry of urban movement is fundamental in human psychology and sociology studies, in the process of developing operationally efficient policy initiatives and management strategies for businesses and the economy. Traditional urban mobility and movement studies have typically relied on stated surveys, which inherently do not accurately reveal behavioural attributes. These stated preference surveys further suffer from low sample ratios as comprehensive samples are often expensive to conduct, making a case for the increasingly readily available typically exhaustive revealed consumer data.

Volumes of novel individual-level spatial-temporal data are now readily available from a variety of ICT and pervasive technologies, and close-circuit televisions (CCTV’s). Examples of this novel information include passenger

[1] BR in this instance is the abbreviation for British Rail

ticketing data, mobile phone Global Positioning System (GPS) presence used to capture part of passenger's origin to destination journeys, and CCTV imagery within and outside a railway infrastructure. These consumer information contain a wealth of unexplored data about peoples' behaviour, with the potential to revolutionise our understanding of social phenomena. These novel data have been called 'big-data'[2] due to their volumes, velocity, wide variety and veracity. The LENNON[3] (ATOC, 2003a) ticketing database records 3.7m daily ticket transactions, representing 756 different ticket product descriptions, sold from over 8000 ticket machines in stations, on trains and over the web, thus forming so called 'big data'. Big Data has been identified in the UK as one of the great scientific and economic drivers for the next decade (Willetts, 2013), and this has spiralled massive research investments in facilities where such datasets could securely reside in readiness for safe dissemination for research purposes.

The strengths that big-data brings includes, high sample rates and variability yielding higher fidelity in reconstructing behaviour and a broader range of solutions (Whittaker, 1935), potentially revealing behavioural activity with accuracy and precision. A drawback of big-data are that it reveals the utility maximising aptitude of humans and may not always contain enough variability to ascertain the full spectrum of behaviour (Adamowicz et al., 1994). An example would be the choice behaviour of consumers given a choice set of 10 products. Most of the consumers will go for the top three products if they represent the highest utility, thereby not providing enough consumer information for the remaining products, thereby exhibiting a limited

[2] Big data is a term coined to describe the range of novel large datasets that are increasingly readily available, like the comprehensive rail ticketing dataset, and other data from sensors of consumers of digital services. Business analysts refer to the five 'V's, and describe big data as having volume, velocity, variety, veracity and value. In GIS, big data refers to data that are large and unstructured that they do not fit on to conventional hardware and software information tools.

[3] LENNON, Latest Earnings Networked Nationally Over Night, is one of the world's largest transport settlement systems capable of handling and allocating to Britain's passenger train operators, more than £3.5 billion in annual passenger revenue from over 450 million train tickets sold via more than 8000 ticket machines in stations, on trains and over the web. Each ticket sale transaction is polled, validated and apportioned by LENNON and assigned to the relevant (ones out of about 27) Train Operating Companies (TOCs) within 24 hours. LENNON is managed by the Rail Settlement Plan (RSP) on behalf of the train operators.

range of sensed attributes. In a stated survey such issues can be better controlled, perhaps by sampling equal amounts for each product category. The unique attributes of novel sources of consumer data make a case for reviewing traditional analysis methods traditionally adopted for measured stated surveys, to integrate the wealth of increasingly available novel big data. Furthermore, traditional geospatial analysis methods have been developed on the assumption that data are measured stated surveys. Big data and consumer data are distinct, as they typically reveal a skewed subset of the wider population who consume the digital service, and as such conventional methodologies developed for traditional measured state survey data need to be reviewed for application to these novel datasets.

This research focuses on case studies in West Yorkshire as shown in Figure 1.1. The inland metropolitan county of West Yorkshire has a population of 2.2 million, and covers 2030km^2. Major railway lines traverse the county, serving about 20million rail passengers monthly. Mobility analysis from the 1950's onwards only had to anticipate aggregate number of passenger trips in deciding policy on transport infrastructures to install. With operational effectiveness needs and saturated urban areas, the travel behaviour and activity patterns of individuals now have to be fully understood. Representative models of the physical environment and urban morphology have to be reconciled with individual-level urban mobility in deciding policy on the operational efficiency of infrastructure.

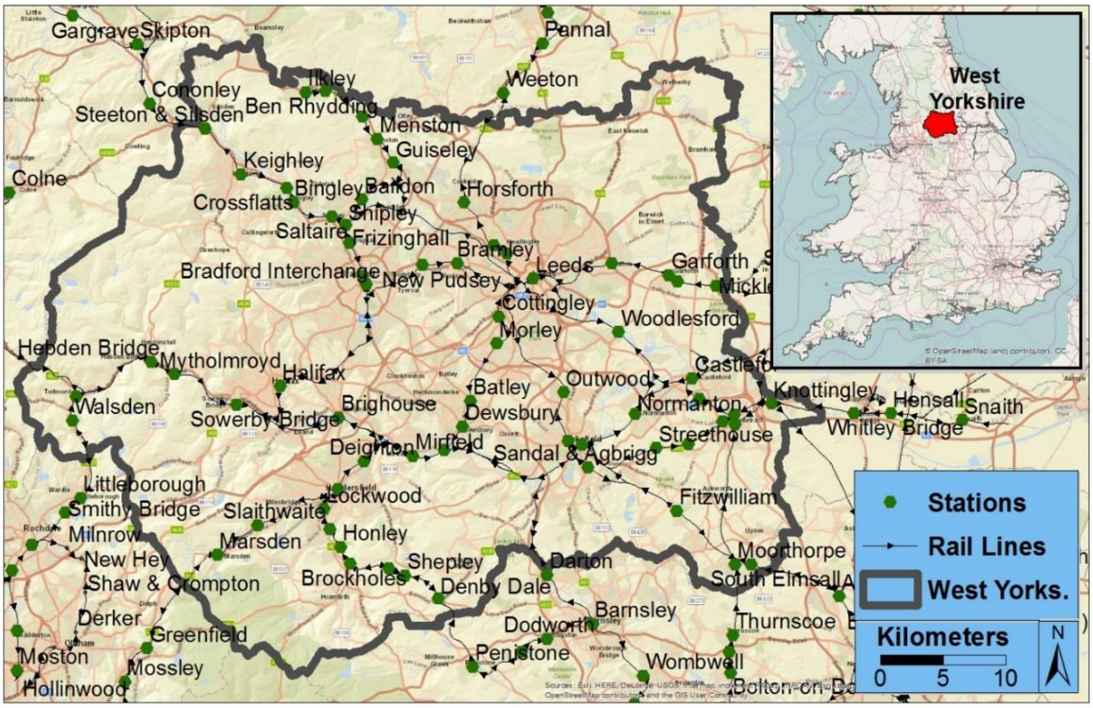


Figure 1.1 Study area of West Yorkshire County UK.

1.2 Urban mobility concept

Urban mobility refers to the ability to move freely and easily, including travel, transport and communications, in time, physical space and in virtual environments (Internet). Urban mobility includes concepts like group mobility, social mobility and migration (Buscher, 2014). In this thesis, urban mobility refers to the movement processes of individuals with a view of describing their activity and decision making processes. Movement patterns are usually established by developing a parametric model of mobility, representing the conceived dynamics (Batty et al., 2012). The rail industry in the UK daily transports over 4m passengers, achieved by about 25 Train Operating Companies (TOCs) through Network Rail infrastructure, which includes a network of about 2500 stations, 21,000 miles of track defining 11,000 route miles (Rail Delivery Group, 2015). Traditional transport analysis is often based on aggregate data on number of trips; however, increasingly (Chapin, 1974, Hägerstrand, 1970) there has been an interest in understanding activity patterns in relation to information models of the environment.

1.2.1 Trip based mobility

The classic trip-based concept of mobility analysis serves the purpose and has benefited from decades of research and intellectual effort to produce the classic 4-stage transport model (Ortúzar S and Willumsen, 2011). Many variants and improvements of the classic 4-stage transport model exist and these have incorporated developments over the years. However, the primary desire to fulfil an activity is what secondarily yields a passenger trip, so in principle an understanding of mobility phenomena will be better achieved by concentrating on the primary activities. The activity-based methods concentrate on patterns of activity and tour models (Bowman and Ben-Akiva, 2001) and have evolved, better incorporating time-space geographic constraints, interactions, change and adaptations in mobility which better explain why people travel and describe the phenomena of individual movement patterns.

1.2.2 Alternative mobility concepts

Activity-based methods evolved from concepts applicable in traditional trip-based transport models (Jones et al., 1983), and in many instances the traditional trip-based and the novel activity-based concepts in mobility analysis complement one-another. As such mobility analysis is potentially better served by seamless integration of trip- and activity-based methods within a framework. The strengths that traditional trip-based methods bring is a wealth of established geospatial and transport modelling techniques. These are suitable for empirical validation using stated preference survey data and for potentially exploiting opportunities of using novel revealed preference type consumer big data. The more recent multi-agent paradigm combines trip and activity based modelling concepts of mobility by creating attribute rich micro-level demand data, useful to both trip and activity genres. In many practical situations, the choice of mobility modelling concept or method (trip, activity, multi-agent) has largely been decided by the type of data available for analysis (Dong et al., 2006, Horni et al., 2016).

1.3 Research questions and hypothesis

A summary of the research questions addressed in this thesis are as follows:

- Would the pre-processing stages of consumer data have an impact on its usefulness?
- Can railway consumer data be harnessed to augment estimates of passenger attributes and activity?
- Can the context of ticket use be identified from a logistical network model?
- Would current industry infilling methods be improved by developing a principles methodology, exploiting micro-level ticketing data?

The details of the research questions and the hypothesis are as follows: The first research question asks whether the rail sector ticketing data pre-processing stages have a significant impact on its usefulness in identifying facets of passenger mobility, and in revealing the mechanism through which the data was created and became available. The hypothesis is that the mechanism through which the missing values occurred in the ticketing data can be established by unravelling the LENNON ticket journeys into time-series cross-sections, separating the outward and inward parts of the journey, establishing factors and offsets for the daily rates of ticket[4] use (journey factors) and ticket validity periods, and creating augmented data for flows to so called group stations.

[4] The daily rates of one-way and return tickets are assumed known and are set at 1, and 2 respectively. However, the daily rates of season tickets are a behavioural attribute peculiar to an individual passenger. A 7-day season ticket can be used 10 times by a regular commuter, and 14 times by a salesperson visiting multiple locations daily.

The second research question is: can big consumer data generated within the rail sector industry be harnessed to augment current industry estimates of individual-level passenger attributes and mobility activity? The hypothesis is that such harnessed data would have a significant impact on the quality of inputs to rail industry planning models. Accurate and representative micro-data are significant to network operators in the rail industry, and often are the major impediment to successful strategic, tactical, and operational analysis.

The third research question is: can the context (for example crowding, waiting, transit stops, etc.) in which the ticketing data were generated be identified to aid in causal inferences on drivers of mobility phenomena on the rail network? The hypothesis is that an accurate Geographical Information System (GIS) rail network model with accurate transit schedules information, would create a logistical model of the rail network. If a representative population of railway passengers are used as inputs to such a model, the context of mobility can be identified by overlaying the events associated with each passenger. Further, such a model would be beneficial to network operators for the management, maintenance and interventions assessment on the railways. The GIS would serve as an integrator, analyser and visual thematic display (Egenhofer and Kuhn, 1999), seamlessly synergising the various tools and applications utilized for railways passenger and cargo mobility and infrastructure analysis.

The final research question is: would it be possible to augment current industry infilling strategies for missing values in ticketing data, with a principled methodology based on scientific best practices? The missing values occur in the variables for daily rates of ticket use (journey factors), tenures of flexible validity tickets, and flows to group stations. The hypothesis is that the current industry strategy of using proxy trip rates for season tickets and logical rules based methods to assign passenger journeys made with multi-modal tickets, can be augmented by a principled imputation method. An improvement in the detail of passenger demand through the use of big data will enable the investigation of a broad range of case studies.

The research indicates that Little’s missing test (Little, 1988) can guide the identification of relevant data variables and values required to plug gaps in consumer datasets. On the basis of this, relevant additional datasets are sought for integration with ticketing data. These are socio-demographic passenger attributes (like income, journey purpose etc., typically available within the Census), and information relating to passenger behavior due to service provisions during their within-network journeys (typically available within the NRTS). Once the relevant datasets are identified, it is often found that the process of integration requires adjustments in resolution and in geographies of scale to maintain consistency between the datasets (Deming and Stephan, 1940). Often one dataset has to be systematically adjusted to fit the resolution, and system process (Howe et al., 2008, Weber et al., 2014) associated with other datasets, and the adjustment methodology adopted heavily impinges on the quality of thesis. The indication is that modelling mobility at finer scales requires integrating big datasets with other relevant conventional data sources, and the validation of such integration impinges on the quality of such models.

Rubin’s work (Rubin, 1976) indicates that for objective imputation and analysis of data, the missing data mechanism has to be distinct from the mobility mechanism being inferred, and that any additional relevant variables with sought attributes must explain the difference in observed and missed values. The Bayesian methods in particular have useful application as missing data are imputed within the context of a mobility model, it provides for the inclusion of separate missing data models, and it facilitates easy integration of additional relevant information.

1.4 Aims and objective of thesis

A summary of the aims and objectives of the research project are listed below, followed by further details:

1. Develop a methodology for harnessing consumer datasets available within the UK rail sector to create a requisite dataset, rich enough that they can be deployed to any of the transport modelling concepts (trip, activity, multi-agent).
2. Study individual-level mobility and movement patterns to enable better forecasts of passenger demand.
3. Develop novel techniques based on statistical best practices, for imputing missing values in mobility variables within rail sector consumer data.
4. Create a Bayesian modelling framework for spatial analysis of harnessed rail sector ticketing data, revealing the mechanism of a range of mobility phenomena.

To meet the challenges of efficiently operating the railways while ensuring economic, social and environmental sustainability, innovative ways are needed to analyse the multi-variant inter-relationships between the facets of the network infrastructure, passengers and surrounding habitats, and to visualise these railway network mobility interactions. This research develops such a framework for urban mobility and movement patterns analysis. The research explores opportunities of reviewing modelling tools, using consumer big data available within the UK rail sector to develop new methods for understanding, simulating and predicting individual movement patterns. The aim is to consider frameworks relating population mobility to demographic characteristics, relative location, time of the day, and other likely drivers of behaviour and interaction on the railways.

The key to any urban mobility exercise is in part to establish the volume, characteristics and location of commuters and passengers in the network (Lomax et al., 2013). In the absence of an attribute-rich, comprehensive and representative population, a simulated population is often necessary for the analysis of complex mobility on the railways. In this research, effort is focused on harnessing available large rail sector consumer datasets, to create a requisite dataset, rich enough that they can be deployed to any of the transport modelling concepts (trip, activity, multi-agent). An attribute-rich synthetic population of individuals including both exogenous and endogenous attributes would enable an accurate assignment of the passengers to the rail network, thereby precluding the need for complex transit assignment in traditional transport models. Further, on the strength of Rubin’s work on missing data (Rubin, 1976), spatial analysis strategies are developed for imputing missing behavioural passengers attributes, as well as for analysis of a range of case studies, enabling the detailed mechanism of a range of mobility phenomena to be identified.

A summary of the objectives to fulfil the aims of the project are enumerated below:

1) Identify variables from relevant databases that plug gaps (by explaining the difference between observed and missing values) in railway ticketing data **(Chapter 3)**.
2) Explore a range of deterministic and stochastic constrained optimisation (spatial microsimulation) methodologies, to develop a strategy suitable for combining *skewed* revealed *interaction* data with data from measured stated surveys. This creates a representative

micro-level population rich in exogenous attributes **(Chapter 4 and Chapter 5)**.

3) Build a GIS network dataset constrained by detailed transit schedule information from the General Transit Feed specification (GTFS) database, and feed the individual-level synthetic population through the logistical GIS network. This creates a population rich in endogenous (within-network) attributes, like volumes in train or waiting times each passenger endured **(Chapter 6)**.
4) Create an animation for visualising the individual passengers and the transit trains in mobility within the railway network, with a query facility to identify context of individual mobility and overlaid aggregate mobility. This adds contextual information to the micro-level population **(Chapter 6)**. Visualisations make it easier to assimilate large amounts of passenger activity.
5) Construct a Bayesian mobility and imputation model (**Chapter 7**) which includes the following constructs:
    - Passenger frequency counts modelled as a negative Binomial distribution, ticket validity periods as Poisson distribution, and group station flows as a Bernoulli distribution.
    - Daily rates of ticket use (journey factors) modelled as a mixture of normal distribution.
    - Journey factors truncated to be within industry experience of limits of journey factors (~3)
    - Ticket validity periods censored to be shorter than the maximum ticket validity.
    - Flows to group stations as augmented flows restricted to sum to one flow (using the ‘dsum’ construct).
6) Imputation of the missing values in variables for journey factors, ticket validity periods and group station flows, and perform sensitivity analysis on an intuition about priors of drivers of mobility (**Chapter 7**).
7) Investigate the case study of rail-heading using a Conditionally Autoregressive (CAR) model with a non-structural Besag-York-Mollie (BYM) component (**Chapter 8**).
8) Investigate the intervention (new building infrastructure) of a new rail station using a Geo-statistical kriging model (**Chapter 8**).

1.5 Conceptual methodology

The skew, unobserved variables and missing values in the LENNON ticketing dataset represents gaps which need to be plugged to exploit the full potential of LENNON as a data source for urban mobility. Heckman (1977) highlighted the bias in analysis results due to gaps resulting from ‘omitted variables’ and ‘skewed samples’. The lack of context of use of these ticketing data further presents challenges in causal inference on the drivers of mobility choice. Rubin (1976) proved that if, as in the UK railways, mechanism of missing data values - a design feature of the railway measuring sensor- is distinct from the phenomena of railways mobility being inferred, then the simultaneous behavior (joint distribution) of the mobility variables is simplified. If new variables that explain the difference in the observed and missing data are found and incorporated in the model, then the missing data would be described as ‘MAR’ or missing at random (Little, 1988, Rubin, 1976) with respect to the new dataset, and the previously intractable joint distribution becomes amenable. As such Rubin’s seminal paper on ‘missing values’ (Rubin, 1976) laid the framework for addressing missing data problems today. In this thesis a set of concerted complementary methodologies are developed for identifying other relevant datasets to combine with available rail sector data, to plug the gaps and to create a requisite dataset for subsequent mobility analysis.

Spatial micro simulation techniques are investigated for use in combining rail sector big data (LENNON ticket data), using reference surveys (NRTS[5] and 2011 Census) as a target set, to supplement absent passenger attributes. In this thesis both deterministic and stochastic spatial microsimulation strategies are discussed for combining various datasets, creating a representative micro-mobility population of railway passengers embedded in the wider population within a geographic region (West Yorkshire, UK). A deterministic methodology particularly suited to adjust skewed rail-mobility consumer big data are developed, combining information on all rail tickets sold in the UK, with the 2011 Census commute to work data and the NRTS,

[5] National Rail Travel Survey (NRTS) was conducted in 2004/2005 in the UK, in areas outside London. The NRTS measured passenger characteristics like access and egress modes to railway network, passenger final destination and home address, purpose of travel, ticket type, and time of arrival at station, travel start time, as well as a range of socio-demographic characters.

yielding the representative micro-level synthetic population. The deterministic strategy using multi-dimensional iterative proportional fitting (m-IPF) is presented in a concise and accessible way, highlighting nuances, precautions in practical application, and associated advantages of the methodology. The spatial microsimulation methodology presented in this thesis aims at extending for the first time the concept of simultaneously constraining 3-levels of variables, enabling the inclusion of origin-destination information associated with mobility interaction, in the spatial microsimulation process. The simulated population created includes weights which represent the probability density for rail commuters and passengers, such that a sample of the synthetic population according to the density yields representative rail passengers, whilst a uniform sample yields the wider population sample.

The attribute-rich synthetic population of individuals produced from spatial microsimulation (inheriting attributes from LENNON, NRTS and the Census) are representative of all railway passengers in the study area, and these are used as input into a rail GIS network that has been logistically constrained by real transit (GTFS) schedules. This produces a rich dataset of railway passengers including both exogenous and endogenous attributes, identifying the rich space-time context and volumes of individual passengers on trains, platforms and stations. The objective is that the representativeness and contextual detail of the synthetic passenger demand population will enable an equitable assignment of the passengers to the rail network, precluding the need for the complexities of a utility optimizing transit assignment.

Bayesian strategies are developed for imputing missing values in rail sector ticketing data. The missing data values are information on passenger daily trip rates (journey factors) especially with season tickets, accurate tenures of flexible tickets, and precise flows to grouped stations. For imputation, a non-parsimonious and comprehensive model is a desired feature (Gelman, 2007, Knutti and Sedlácek, 2013), and all the relevant covariates derived from the spatial microsimulation and GIS network simulation are included. This ensures the inclusion of most of the passenger heterogeneity. For causality and interventions case studies however, there is a need to balance the parsimony of the models with a need to capture relevant complexity in urban mobility (Angrist et al., 1996, Ho et al., 2007, Pearl, 2009, Shadish et al., 2002).

1.6 Scope and structure of the thesis

The structure of the work presented is as follows. Chapter 2 introduces the thesis, starting with a preview of transportation concepts, a review of the concept of big data, consumer data and the gaps that occur in such datasets as it introduces bias in the thesis. Spatial microsimulation is reviewed as a tool for combining disparate datasets to create a demand population that are used in mobility modelling. An introduction of traditional trip-based mobility models and the alternative activity-based concepts of mobility are presented. The logistical GIS network models that identify the context of mobility data are presented, as well as a review of the literature on Bayesian methods developed for data imputation and the analysis of urban mobility and movement patterns.

In Chapter 3 the setting and a brief description of the study area are presented, as well as the data pre-processing. The Little's test is presented and reviewed for use in identifying the nature of other relevant datasets that are necessary to combine with ticketing big data, to produce an enhanced dataset. Little's test for investigating the nature of the missing data mechanism is presented. The preliminary data pre-processing are presented as they enable the reconciliation of the geographies and variables in disparate datasets. In Chapter 3 also, a preliminary summary of the ticketing data are presented to enable an initial visualisation of the mobility interaction.

Chapter 4 presents spatial microsimulation, the first of three concerted complementary methodologies for harnessing big data. The ranges of spatial microsimulation strategies are assessed to establish the accuracy of their results under different input data-types. This assesses the suitability of each methodology for application to novel consumer data inputs. The spatial microsimulation assessments are for when the sample ratio of the seed could take on a range of percentages, when the number of variables informing the target population is limited, and the effect this might have on predictive values for the wider population attributes. Finally, as consumer data tend to be a skewed representation of the wider population, the effect of skewed seed data on the spatial microsimulation is assessed.

In Chapter 5, the spatial microsimulation methodology is used to enhance the railway ticketing data for the West Yorkshire study area. For expedience, the flows analysed are restricted to those that are initiated and terminated within the West Yorkshire study area. The NRTS is considered a skewed dataset as it represents a subset of the population who travel by rail. The

LENNON ticketing data are also skewed as they represent only the population of railway users, whilst the Census interaction data are 100% sample ratio of the UK population. The multi-dimensional iterative proportional fitting algorithm is developed for application to these skewed input data. A 3-level hierarchical target constraint is developed to account for the origin-destination interaction inherent in mobility phenomena. This produces a representative population of railways passengers, embedded in the wider population, and inheriting the attributes of the datasets involved.

Chapter 6 presents details of the second concerted complementary methodology developed to harness consumer data available from the rail sector. This is GIS-GTFS network modelling. In the method, the procedure for incorporation of the general transit feed specification (GTFS) information into a GIS network dataset is detailed in a practise oriented way. The resulting model is called the GIS-GTFS network simulation model, and passengers that are fed into this logistical network were derived from the earlier spatial microsimulation procedure. Each passenger’s times of arrival to the train station and time of first train facilitates a solution of their route on the GIS-GTFS network. The details of events on the network are overlaid, and endogenous attributes are extracted for each passenger, forming an additional rich set of attributes.

Chapter 7 presents the construction of a fully Bayesian predictive model of mobility on the railways. For prediction, the pertinent covariates derived from the spatial microsimulation and GIS-GTFS network simulation are included in the Bayesian model. In Chapter 7, imputation strategies in the literature are reviewed, highlighting the advantages of the Bayesian imputation method, leading on to the choice of such strategy. Details of the directed acyclic graphs (DAGs) which form the basis of solutions preferred for the Bayesian models are presented. The Markov chain Monte Carlo (MCMC) algorithm of the BUGS software (Lunn et al., 2012) are used to implement the Bayesian model. The strategies are presented for the censoring, truncation, and ‘dsum’ constructs that enable restrictions in values of journey factors, ticket validity periods, and augmented flows to group stations. Results are presented for the imputed missing values.

The case studies are presented in Chapter 8. These pertain to analysing the mechanism of rail-heading within the study area. A Bayesian model is created, which adjusts for various covariates, and estimates posterior influences on rail-heading that are due to within Postcode and between Postcode effects. The Bayesian model estimates the proportion of rail-heading that are due to the covariates that were adjusted for. Transit flows that do not originate and terminate in the West Yorkshire study area are excluded, thereby precluding the analysis of any rail-heading associated with and occurring at West Yorkshire boundaries. The effects of a new train station at Kirkstall Forge are also investigated in Chapter 8. A Geo-statistical kriging model is developed whereby observations at train stations are treated as sampling distributions from a continuous mobility interaction across the West Yorkshire County. The details of the projected passenger demand resulting for the new station at Kirkstall Forge are presented and discussed.

Chapter 9 is a discussion and synopsis of the thesis along with projections for application of research finding and its relevance to the rail sector. The overriding importance of ethical consideration in consumer data applications and research are discussed. The conclusions of the thesis are also presented in Chapter 9, as well as the limitations of the work. The work anticipated for the future are also presented in Chapter 9.

The thesis does not strictly follow the IMRAD (Docherty and Smith, 1999, Sollaci and Pereira, 2004) structure of scientific presentations (i.e. introduction, literature review, methodology, case study and results, discussion and conclusion). Instead, in Part 1 of the thesis, an introduction in Chapter 1 is followed by a short literature review in Chapter 2, serving to overview and introduce the concepts developed in the thesis. The brief literature review is followed in Chapter 3 by setting the scene, introducing the study area, the analysis dataset used, and some conventional data pre-processing. In Part 2 the set of three concerted methodologies which form the crux of the thesis are presented separately in individual chapters, highlighting how they complement one another. In each of these chapters (Chapters 4, 5, 6, and 7), a more detailed literature review is presented, as well as methodology, results, discussion and remarks. The chapters in Part 2 are presented in this way because the methodologies are multi-disciplinarily, divergent (as each method deals with a different type of data gap), and the presentation would read less succinctly if they are discussed jointly. In Part 3 (Chapter 8, and 9), the thesis reverts to the traditional IMRAD format.

Some of the computer models and codes used to generate the results have been included in the Appendices. These highlight the pertinent data pre-processing stages and enable the results reported in this thesis to be reproduced. The main datasets are the 2011 Census, NRTS, GTFS, and LENNON ticketing data. Some of these datasets are publicly available, and others can be requested from ATOC.

1.7 Summary table

Thesis aspect	**Brief description**
Aims	The aim is to explore opportunities of integrating consumer dataset within the railways with other relevant datasets, to construct a more detailed picture of passenger numbers, journeys and to relate these through spatial analysis to other drivers of mobility.
Objectives	1. Identify variables from relevant databases that plug gaps in railway ticketing data **(Chapter 3)**. 2. Explore a range of strategies for combining revealed skewed interaction data with data from stated surveys. **(Chapter 4 and Chapter 5)**. 3. Build a GIS network dataset constrained by detailed GTFS transit schedule information. **(Chapter 6)**. 4. Create an animation for visualising the individual passengers and the transit trains on the railway network. **(Chapter 6)**. 5. Construct a Bayesian analysis and imputation model, to study individual-level mobility (**Chapter 7**). 6. Investigate the case study of rail-heading and the intervention of a new rail station (**Chapter 8**).
Research questions	▪ Would the pre-processing stages of consumer data have an impact on its usefulness? ▪ Can railway consumer data be harnessed to augment estimates of passenger attributes and activity? ▪ Can the context of ticket use be identified from a logistical network model? ▪ Would current industry infilling methods be improved by developing a principles methodology?
Summary of methods	1. Spatial microsimulation to combine relatively skewed and disparate datasets 2. GIS-GTFS network model to simulate context of passenger mobility 3. Bayesian imputation and analysis, which incorporates additional passenger behavioural information and rail sector knowledge

Chapter 2
REVIEW OF CONSUMER MOBILITY ANALYSIS

In this thesis, a framework for the use of large consumer data (big data) in urban mobility and movement pattern analysis is developed. This chapter is a literature review of the methodologies used to improve the consumer ticketing data. The concepts of the methods are highlighted to show how they are concerted. The ticketing data are from the UK rail sector, and after improvement they are used in transport modelling and mobility analysis. The review explains how the separate methodologies fit into the wider project. Traditional transport models are first reviewed, because new transport modelling genres that are developed for novel consumer datasets, have evolved from the traditional methods. Then, a review that defines and critiques consumer data, using the example of rail sector ticketing data is presented, and the definitions given in the literature to describe the nature of the gaps in consumer data are presented. The strategies for identifying relevant variables used to plug the gaps in consumer data are reviewed.

The chapter describes what is meant by spatial microsimulation and how it is used to combine variables in disparate datasets. Spatial microsimulation methods are briefly reviewed, as they are used to create a synthetic attribute-rich micro-level demand population that is traditionally used as input to transport models.

The range of trip and activity based transport modelling genres are reviewed, along with the more recent multi-agent modelling genre. The review forms a basis for highlighting areas of potential improvement. The important concept of using GIS to further improve the synthetic demand population used in transport models is presented. The recent availability of detailed transit schedules (GTFS) used to complement GIS network models is reviewed, as these help to create a contextually rich micro-level demand population. Data imputation strategies used to plug missing values in datasets are reviewed. The Bayesian modelling is introduced as a framework for simultaneous imputation (to identify idiosyncratic passenger behaviour) and analysis of micro-level data. Literature on the concepts applicable in the development of Bayesian models are reviewed, highlighting its usefulness in spatial analysis of a range of urban mobility case studies. The chapter ends by presenting an overview flowchart of the project.

2.1 Review of the transport concepts

The traditional transport modelling paradigm for mobility analysis is to establish the volume, distribution, location, and perhaps characteristics of passengers on the network. Traditional transport modelling techniques will predict the volumes generated-by and attracted-to a zone using empirical estimates based on the size, morphology and geodemographic attributes of the zone. The propensity and potential of a zone in such terms determines the volumes of zonal passenger origins and destinations. These strategies enable an estimation of the passenger demand for use in transport models (Ortuzar and Willumsen, 2002). In other transport applications, a count at specific traffic junctions (Willumsen, 1981) or a population survey (Richardson et al., 1995) will give an indication of the demand for use in the transport models.

In modern applications the demand can be estimated from the trace revealed by passengers when they are in the process of consuming digital services while fulfilling daily activities. Developments in technology have meant that it is now more readily possible to sense this revealed data. In the particular case of the rail sector, the consumer data depicting the volume of passengers are the ticketing data which can potentially be exhaustive. However, there are issues with consumer data which prevent their direct use, and these represent the so called gaps in novel consumer data. As discussed in Chapter 1, these gaps come in the form of an incomplete set of covariates related to the attributes of passengers who might have used a ticket (like the passenger journey purpose, or socio-demographic attributes), skew in the distribution of these consumer data relative to a wider representative population (like relative to the Census), and lack of observation of the context of mobility (like the crowding in the train when the ticket was used or the waiting times a passenger might have endured). Further gaps are the missing values about passenger daily rates of season ticket use, number of days of use of flexible period tickets, and precise flows to group stations (SDG, 2014, SDG, 2016), and flows associated with regional multi-modal tickets. (This latter multi-modal so called PTE flows are not dealt with in this thesis[6]).

[6] There is a wider point here that the definition of missing data really depend on what the analyst wants to do with the data /resulting model. For some types of modelling, particular data are not needed and so are not considered missing.

The constrained optimisation method called spatial microsimulation is traditionally used in the field of Geography (Deming and Stephan, 1940) to facilitate taking a survey sample and replicating it in an optimal way such as to create a micro-level representation and the volume of a wider reference population like the Census. For transport applications, this technique is used to create a micro-level synthetic population used as the demand for input to transport models. Different types of transport models serve different purposes and require particular demand inputs, depending on whether the unit of mobility in the transport model is the trip a passenger makes, or the activity patterns a passenger creates. As the particular transport model depends on whether trips or activities are the nomenclature, a review of trip and activity based mobility concepts provides a broad picture of transport modelling concepts. Multi-agent transport models (Rieser et al., 2007) have the potential to combine the trip and activity based concepts yielding multi-agents with attributes relating to both individual trips and individual activities.

In conventional transport models (gravity models, direct demand models, multi-agent models, etc.), having established the origin and destination demand, it is typically possible to estimate the passenger modes of transport and route choices using methods established in the literature (de Dios Ortuzar and Willumsen, 1994, Hensher and Button, 2007). However, it is particularly difficult to ascertain the context of passenger mobility. Such information would be pertinent in causal inference on passenger behaviour and on drivers of mobility. The difficulty in establishing context of mobility is a result of a dearth of sensors on all modes of transport to measure the idiosyncrasies of passenger behaviour relating to the choices they make. Individual confidentiality issues also limit the amount of information accessible from GPS traces of passenger mobility. In the case of the rail sector however, detailed information on transit schedules (GTFS) are now available to the public. The implication is that a logistical rail network can be built by incorporating the GTFS information within a Geographical Information System (GIS) (Longley, 2005, Poletti et al., 2017). Assuming then it is possible to synthesize a micro-level demand population of rail passengers with attributes on passenger arrival times at train stations, first trains taken, number of stops, etc. In such a scenario, if the representative population are fed through a representative logistical GIS network, the overlaid events would represent the context of passengers experience on the rail network. This would yield contextual information, like the crowding in the train when each passenger made their journey or the waiting times each passenger might have endured. These concepts are developed in the thesis, facilitated by developments in technology that have created readily available individual-level large consumer data. These new developments form the basis of new transport modelling genres. As such, novel sources of large consumer data have the potential to alter traditional transport and multi-agent modelling genres by for instance precluding the need for complex route choice and utility maximizing traffic assignment in transport models (Horni et al., 2016, Ortuzar and Willumsen, 2002). These issues are explored in detail in this thesis.

The Bayesian modelling framework (Albert, 2009, Gelman et al., 2014b, Kruschke, 2014) which is becoming increasingly popular for the analysis of complex phenomena is presented in this thesis, emphasizing its relevance in mobility analysis. In situations where it is required to construct a detailed picture of mobility, the missing values like the passenger daily rate of use of season tickets (Journey Factors) can be objectively identified using Bayesian models. Bayesian models enable the imputation of such missing values to be accomplished within a mobility model, thereby better reflecting the mechanism through which the data was created in the first place. The additional advantage of the Bayesian framework is that it consists of simple constructs (Lunn et al., 2012, Plummer, 2003) that facilitate parsimonious modelling of complex phenomena. For instance the Bayesian truncation and censoring construct can be used to good effect in mobility models to reflect that although the journey factors are not known completely, however, from local surveys the rail industry has empirical knowledge of its range of values dependent on the length of the journey and time table restrictions. In the Bayesian framework, a mobility model can be informed of such empirical knowledge using simple constructs (Lunn et al., 2012, Plummer, 2003, Turnbull, 1976).

2.2 Consumer data and the rail sector

As briefly discussed in Chapter 1 consumer ticketing data are now increasingly available in the rail sector. However, traditional datasets for railway mobility analysis have been measured stated surveys, designed to be random samples representative of the entire population. Typical examples relevant to the rail sector would be the National Rail Travel Survey (DfT, 2013a), and the Census interaction data (Stillwell, 2006), the latter being a 100% sample ratio. These survey data would typically be conducted within the railways or at residences within the wider populace. Typically, a range of passenger socio-demographic attributes, journey origin and final destinations, and times of rail station access, first train, egress etc. are stated and enumerated. Developments in technology have however, meant that revealed data can be sensed more readily, and this has resulted in increasing available novel large consumer datasets (Kitchin, 2014, Manyika et al., 2011, Marr, 2015, Viktor and Kenneth, 2013). A good example of such data is the rail sector ticketing data from ATOCs LENNON database. These ticketing data can be procured from 8000 points of sale, with about 4.7m daily tickets sales, and about 756 different ticket products representing the rich heterogeneity in passenger use.

The rail sector ticketing data was originally meant to sum revenues for allocation to regional train operating companies (TOCs), but a new use for mobility analysis has meant that accurate and precise flows on the network, associated with each ticket, now needs to be known (SDG, 2014, SDG, 2016). Each ticket further needs to be associated with a passenger to relate socio-demographic attributes to drivers of mobility. As a result, the number of journeys per ticket (journey factor) now needs to be ascertained. Currently, numbers of journeys made with season tickets are unknown. A single or return ticket would be used once or twice respectively, but a 7-day season ticket for instance would have a range of daily rates of use. A lecturer using such a ticket for commute to work during the week would perhaps use such a ticket about 10 times, i.e. an average daily rate of less than twice. A student using such a ticket to get to early and late lectures within the 7-day validity of the ticket would perhaps use the ticket more than the twice daily journey rate. Further, some train stations are grouped together so that passenger flows to specific stations within the group are unknown. Such requisite information is necessary to build an accurate picture of mobility on the railways. Regional multimodal transport operators sell tickets of specific coverage that are valid for use on buses and trains. There exists no accurate figure on number of rail journeys that are made with these PTE[7] tickets.

[7] Passenger Transport Executives (PTE's) manage multi-modal ticket sales in regional conurbations outside Greater London. These particular tickets were not available to the project, and as such were not analysed during the research. However, mobility associated with PTE tickets are conceptually similar to mobility related to group stations. PTE ticket use are typically restricted to the number of stations within a pre-defined rail-zone, just like group station ticket use are restricted to the number of stations within the pre-defined group. As a result, analysis methods developed for group stations are applicable to PTE tickets when they become available.

The literature (Viktor and Kenneth, 2013) has reported that the attraction of large consumer data lies in the potential contextual richness of such data. In many consumer data scenarios however, the context of the data is not explicitly recorded and this therefore is unknown. In the case of ticketing data the context of the data would represent the passenger's detailed experience on the rail network. This would include how many people were in each train (to represent crowding), passenger waiting times, number of stops, transfers, etc., influencing the passenger's attitude towards the railways. The crowding or long waiting times on trains on particular routes could result in passenger preference for an alternative longer route, with such causal inference only identifiable from knowledge of the context and sequence of mobility phenomena. Additional issues with the LENNON ticketing data, typical of many large consumer datasets are that they represent a skewed subset of the wider population, instead of a random subset because the data has arisen from a specific purpose, not a generalised data collection effort. This poses a challenge in developing methodologies for reconciling or combining these datasets with reference datasets like the Census, which typically contain the sought socio-demographic attributes.

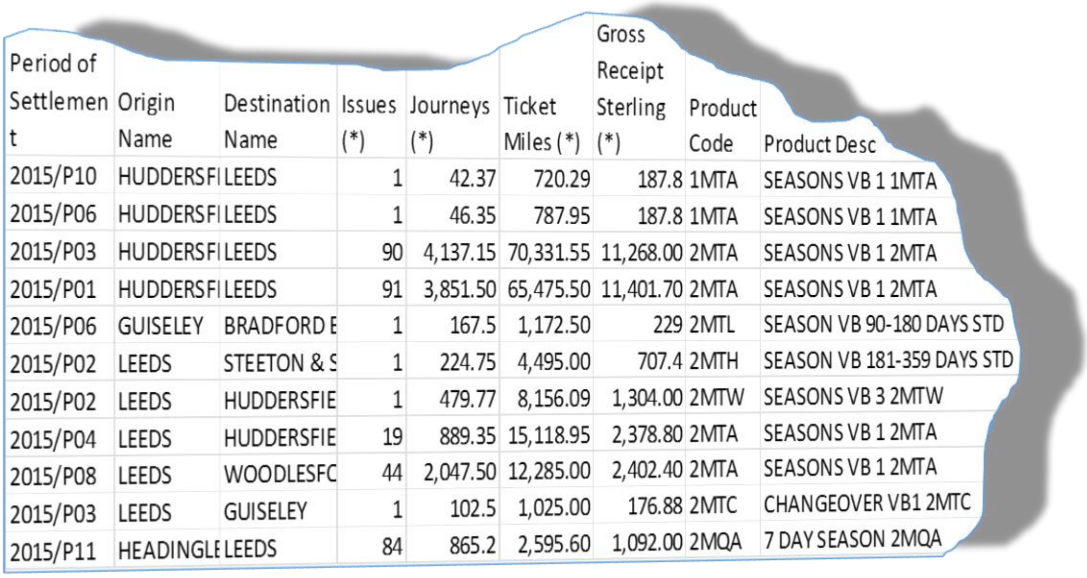
In this thesis, methodologies are developed for addressing the above mentioned gaps and shortfalls of novel consumer data, in particular those available and applicable to the UK rail sector consumer ticketing dataset. Once the data gaps are resolved, the ticketing data is improved, forming a requisite dataset for mobility spatial analysis, thereby enabling a used beyond the original intention for revenue accretion.

2.3 Nomenclature for gaps in data

The different types of gaps in the ticketing data have to be put into context formally to enable objective methodologies to be developed to plug such gaps, thereby improving the consumer ticketing data. Traditionally, datasets are arranged in rectangular attribute tables of records (rows) representing tuples and fields (columns) representing variables as illustrated in the snippet LENNON dataset in Table 2.1. Each row in the table represents a ticket sale to an individual (or to a group of individuals each procuring a ticket product with the same attributes). Each ticket includes attributes like purchase time aggregated into monthly periods for anonymization), station entry (i.e. Origin Name) and station exit, journeys (number of journeys made per ticket issue), estimated track miles covered per ticket and cost, ticket type and additional product codes describing further details on the ticket.

Missing data can occur at a primary level when one or more complete records (rows) are missing and this is called unit omission. At a secondary level when values for one or more variables within a unit are missing it is then called item omission. The LENNON ticketing dataset is assumed exhaustive (i.e. comprehensive) with entire UK coverage, as such this research does not account for infilling methods for the so called unit omission whereby whole tickets are missing. It is assumed that all ticket sales are known and accounted for, but just that some details (values) relating to the number of journeys, specific origins and destinations etc. are missing and require infilling. Cases of unit omissions (Schafer and Graham, 2002) are typically more relevant to stated preference surveys which are designed to be a subset representative of the population.

Table 2.1 LENNON table snippet showing season ticket infills.



Period of Settlement	Origin Name	Destination Name	Issues (*)	Journeys (*)	Ticket Miles (*)	Gross Receipt Sterling (*)	Product Code	Product Desc
2015/P10	HUDDERSFI	LEEDS	1	42.37	720.29	187.8	1MTA	SEASONS VB 1 1MTA
2015/P06	HUDDERSFI	LEEDS	1	46.35	787.95	187.8	1MTA	SEASONS VB 1 1MTA
2015/P03	HUDDERSFI	LEEDS	90	4,137.15	70,331.55	11,268.00	2MTA	SEASONS VB 1 2MTA
2015/P01	HUDDERSFI	LEEDS	91	3,851.50	65,475.50	11,401.70	2MTA	SEASONS VB 1 2MTA
2015/P06	GUISELEY	BRADFORD E	1	167.5	1,172.50	229	2MTL	SEASON VB 90-180 DAYS STD
2015/P02	LEEDS	STEETON & S	1	224.75	4,495.00	707.4	2MTH	SEASON VB 181-359 DAYS STD
2015/P02	LEEDS	HUDDERSFIE	1	479.77	8,156.09	1,304.00	2MTW	SEASONS VB 3 2MTW
2015/P04	LEEDS	HUDDERSFIE	19	889.35	15,118.95	2,378.80	2MTA	SEASONS VB 1 2MTA
2015/P08	LEEDS	WOODLESFC	44	2,047.50	12,285.00	2,402.40	2MTA	SEASONS VB 1 2MTA
2015/P03	LEEDS	GUISELEY	1	102.5	1,025.00	176.88	2MTC	CHANGEOVER VB1 2MTC
2015/P11	HEADINGLE	LEEDS	84	865.2	2,595.60	1,092.00	2MQA	7 DAY SEASON 2MQA

2.3.1 Statistical missing data mechanism

Here an illustrative example is used to explain the statistical mechanisms of missing data (Allison, 2001, Little and Rubin, 2014), using the concepts of missing at random (MAR), missing not-at-random (MNAR), and missing completely at random (MCAR). Consider passenger behaviour. It is known that work activity typically takes place at a specific spatial location and time. In order to model a daily trip (or tour) undertaken with a monthly season ticket between an entry and exit station, the number of times the ticket is used daily could depend on the socio-economic characteristics of a passenger (job type, stage in life, affluence etc.), urban morphology (transport, work, housing infrastructure), other endogenous and exogenous network influences (station access/egress, cost of travel, comfort etc.).

Consider the hypothetical fictitious case of a salesperson and a lecturer. The behaviour of a salesperson for instance who chooses to use the season ticket to travel between sales-work destinations many times a day, would be different to that of a lecturer who commutes to work daily on a season ticket. If the dataset available for mobility analysis included a job-type variable, then it would be likely that similar salespersons who use the train for work-related travel activity would record higher journey factors across season ticket product ranges, and an observation of the job-type variable would capture this behaviour. Within the group of salespersons, if journey factor information on a number of them are missing, it can in principle be deduced or inferred by conceiving a random variation in their behaviour (their journey factors). Within the group of similar salespersons, the frequency of use of season tickets can be treated as a normal random phenomenon conditional on the other attributes of the salespersons. This embodies the concept that any information missing therein is missing at random (MAR) (Little and Rubin, 2014), and objective inference on such missing data can in principle be achieved from just the observed salespersons.

Within a wider group of salespersons and lecturers, had the information on job-type not been available, then a model of missing journey factor information cannot assume a random variation in journey factors, as this variation would more probably be systematic. Under such conditions any missing information is missing not at random (MNAR) (Little and Rubin, 2014, Rubin, 1976), and any objective inference would be based on both the observed information on salespersons and lecturers, and importantly on the unobserved job-type information. This random dependence of season ticket journey factor solely on the observed job-type descriptions in the MAR case, and on the unobserved job-type information in the MNAR case is central in statistical modelling of passenger behaviour and inference on any missing information. MAR and MNAR are terms used to describe the statistical mechanism through which data variables or values became missing. The terminology and nomenclature adopted in the literature to describe this missing phenomena is the '***missingness***' (Rubin, 2004).

A special MAR scenario occurs when the unobserved component of a dataset is the result of a traditional measured random sample of the population. In such a case, objective statistical inference can be drawn by analysing just the sample. Inference in such a scenario will be independent of the particular observed sample, as expectedly an alternative measured random sample on the same population would yield the same results. In addition to this, the inference also does not depend on the unobserved part of the population that inadvertently did not make it into the random sample. This special '***missingness'*** MAR scenario, independent of the observed and unobserved data, is described as missing completely at random (MCAR) (Heitjan and Basu, 1996, Little, 1988, Schafer and Graham, 2002).

2.3.2 Ascertain missingness mechanisms

Many data analysis and imputation strategies are based on the data being MCAR or at the barest MAR. To ensure the validity of the proffered solutions, methods have to be developed to test and ascertain the veracity of the assumptions. However, inferring the mechanism of missing data is not straightforward as there are no tests for this in the literature (Potthoff et al., 2006). Typical ***missingness*** tests developed in the literature rely on some assumptions about the data that in turn cannot be tested (Rhoads, 2012). A missing data test proposed by Little (1988), and applicable to data with a monotone pattern of missing data, ascertains the correlation between sets of covariates to infer dependence. If dependence is established (between a complete case and a covariate with missing values), then it becomes plausible to infer that the variables would complement each other, enabling the missing values to be described as MAR.

In the Little's test, the null hypothesis is that the missing data is MCAR, implying that a p-value greater than 0.05 (i.e. $p \geq 0.05$) is indicative that there is weak evidence against the null hypothesis that the data are MCAR, and the null hypothesis that the data are missing completely at random is not rejected. If the test indicates that $p < 0.05$, then the null hypothesis that the missing data is MCAR is rejected. Rubin (1976) postulated further to ascertain the missing data mechanism. In this case, Rubin asserted that if (as in the rail sector), the missing data mechanism (a design feature of the railway measuring sensor) is distinct from the phenomena (of railways mobility) being inferred, then the simultaneous behaviour of the (mobility) variables is simplified. If additionally new variables that explain the difference in the observed and missing data are found and incorporated in the model, then the missing data would be described as 'MAR' or missing at random (Little, 1988, Odiari, 2016, Rubin, 1976) with respect to the new dataset. Under such conditions, the previously intractable joint distribution becomes amenably simplified. A formal and detailed presentation of this assertion is presented in Chapter 3 (Section 3.2.1), justifying when inference on datasets can be based on just the observed data.

Understanding the missing data mechanism requires that the analyst comprehends the details of the data creation process. Once the data creation mechanism can be traced, it becomes possible to separate the observed data into components associated with each variable crystallisation stage. The process of unravelling the data in this way helps provide an insight and understanding of the missing data mechanism, and this strategy has applications for causal inference (Leung, 2004, Pearl, 2009, Pearl et al., 2016). In the case of the ticketing data which has missing values in the variables for journey factors, ticket validity periods and flows to group stations, the process of unravelling the ticketing data separates the outward and inward flows for return tickets, and creates offset and exposure variables to account for different journeys per ticket and tenures of different ticket types. Additional augmented data are created to represent the range of possible flows to group stations (and a similar treatment could be adopted for PTE ticket use had they been available to the project). These processes highlight and expose the ***missingness*** in the dataset, enabling Little’s test to be applied to indicate the nature of additional relevant variables that would make the dataset MAR. The MAR dataset in turn enables infilling methodologies to be developed based on Rubin’s postulations.

2.4 Review of spatial-microsimulation

Apart from journey factors, ticket validity periods, and group station flows which are variables with inherently missing values within the LENNON ticketing data, the LENNON data has a range of additional attributes (ticket type, origin-destination station, cost of ticket, etc.) defined on the ticket product. For mobility analysis however, a range of additional variables are necessary (like the passenger associated with each ticket, journey purpose, socio-demographic attributes etc.). It is unlikely that all these variables would be available in a single dataset, so methods typically have to be developed to combine available data with other datasets containing the sought additional attributes. When large novel consumer data are considered for use beyond their original purpose, data ownership, legal, and ethical issues are first resolved. Then, a typical data analytics paradigm involves investigating an appropriate theoretical context, then linking to additional datasets with sought attributes to enable the developments of such framework. When a full repertory of attributes becomes available, the data analytics process typically involves an iterative cycle of classification of the variables and data values, regression and predictive analytics.

Due to the disparate nature of datasets, linking is usually more complicated than a database operation or logistic rules on common attributes. Typically the datasets have different resolutions and geographic spatial scales, requiring more sophisticated reconciliation methods. A sophisticated strategy for combining datasets is the spatial microsimulation methodology, developed by population geographers for micro-level population synthesis. Early Census requirements meant that despite complete counts of the population, only a small sample of anonymized records (SAR) of cross-tabulations of all individual attributes were released, alongside comprehensive sets of aggregate tables typically limited to only three variables per table (Williamson et al., 1993). Despite the anonymization and aggregation in respect of individual privacy, population geographers realize that comprehensive and disaggregated records would enable the construction of more detailed pictures of geographic phenomena. This led to a need to reconcile the small SAR cross-tabulation (called the sample or at times proposal distribution) with aggregate Census data (called the target distribution) to create a comprehensive micro-data using the spatial microsimulation methodology (Deming and Stephan, 1940).

2.4.1 Deterministic spatial microsimulation

The first application of spatial microsimulation (also called population synthesis) is widely attributed to Deming and Stephan (1940), who applied the method to the USA Census. A synthetic population was formed from a combination of multiples of individuals in a sample of anonymized records (the seed data) that will best fit the aggregate values defined in Census tables, using the Lagrange multiplier constrained optimization method (Bertsekas, 2014). With the development of computers and numerical analysis, instead of solving the linear equations using the Lagrange multiplier method, iterative methods gradually converging towards an optimum were proposed (Fletcher, 2013, Kelley, 1999). It became increasingly more efficient to resort to these iterative proportional fitting (IPF) strategies, as they are faster to compute, less sensitive to numerical and round-off errors and simpler in algebra, than comparative formulations by Lagrange and Fermat (Fermat, 1891, Lagrange, 1867). The IPF algorithm has found applications in diverse fields such as in geography (Birkin and Clarke, 1988, Harland et al., 2012, Upton, 1985), transport engineering (Duguay et al., 1976, Fratar, 1954, Furness, 1965), economics (Bacharach, 1970), statistics and computer science (Lavrakas, 2008) under various names.

2.4.2 Stochastic spatial microsimulation

Apart from the above deterministic (IPF) spatial microsimulation strategies which yield the same result on repeat, the alternative simulation strategies are stochastic, typically Markov chain Monte Carlo (MCMC) based methods. The MCMC methods are efficient for converging to solutions of intractable complex constrained optimization problems (Asmussen and Glynn, 2007, Brooks et al., 2011). Population geographers have developed a range of MCMC variants to complement traditional deterministic IPF algorithms. Some of these methods have been branded hill-climbing and have been compared to IPF methods by Kurban et al. (2011). The strategies branded simulated annealing have been compared with IPF by Harland et al. (2012). Another MCMC based strategy the genetic algorithm was used to simulate network traffic by reconciling observed and estimated flows, opening the realm for application to transport problems (Dimitriou et al., 2006). Whilst established stochastic MCMC methods are based around the Metropolis Hastings algorithm (Chib and Greenberg, 1995), more recent implementations built around the Gibbs Sampling algorithm have emerged for application to transport (Chib, 1995, Farooq and Bierlaire, 2017). This latter Gibbs based methods are in their infancy of development. The synthetic reconstruction is another spatial microsimulation strategy developed in the UK (Birkin and Clarke, 1988, Birkin et al., 2006), exploiting the probabilistic indicative potentials of the Sample of Anonymized Records (SARs) dataset from the Office for National Statistics (ONS).

2.4.3 Remarks

The difference in the alternative deterministic and stochastic spatial microsimulation strategies lies in the numerical algorithm or probabilistic method adopted for estimating or calculating the weights assigned to individuals in the seed data (to replicate the seed to form the synthetic population). In deterministic methods the weights are arithmetic fractions of items in the proposal distribution, in turn iteratively improved to fit the aggregate target constraints. In stochastic methods, random samples are taken recursively with replacement from the seed data to improve the objective function being optimized until the proposal distribution converges to a target distribution. As a result, deterministic strategies yield fractions of individuals, while stochastic procedures yield integer multiple counts of the individuals. Widely used deterministic strategies are reported in literature (Barthelemy et al., 2016, Lovelace and Dumont, 2016), as well as stochastic strategies (Kavroudakis, 2015, Williamson et al., 1998).

Spatial microsimulation has been reviewed as a useful tool to create a micro-level demand population for use in subsequent transport modelling stages. As the usefulness of the demand population created is dependent on the particular transport modelling genre adopted, the traditional trip-based and activity-based transport models, and more recently multi-agent models are briefly reviewed (Horni et al., 2016, Stillwell, 2006, Stillwell and Duke-Williams, 2003). In the next section, traditional transport modelling genres are reviewed, leading on to details of how novel Geographical Information Systems (GIS) logistical models can further improve the contextual detail in synthetic demand populations, in the process creating a dataset rich in exogenous (outside-network) and endogenous (within-network) attributes, precluding the need for the route choice and transit assignment stages of the traditional 4-stage transport models. The indication therefore is that the availability of novel consumer data has the potential to alter transport modelling genres. This forms the crux of work related to the spatial microsimulation and GIS network simulation presented in this thesis.

2.5 Transport modelling genres

Analytical transport planning models are used to manage phenomena on the railways (Assad, 1980). Historically these have been classed into three model types, reflecting their use in the hierarchy of complexities of railways management (Anthony, 1965). Strategic models are designed to address long-term management issues like building a new station or high speed line on the National Rail infrastructure (Preston, 2012, Treasury, 2011). Tactical models on the other hand address medium-term rail network supply and demand management by simulating patterns of carriage, freight and passenger flows, thereby assigning demand to trains to enable optimal scheduling. Finally, operational models address the day-to-day management activities like timetable setting and re-scheduling. However, a severe impediment to the effective use of these strategic, tactical and operational models is the unavailability of objective detailed seed data on passenger counts and generalized costs (Amos et al., 2010, Dempster, 2012, Leung, 2004, Rao and Rao, 2009), and this research aims to plug that particular gap, by creating a rich requisite micro-level synthetic population, likely to have a significant impact on the quality of inputs to strategic, tactical and operational rail-sector analysis planning models.

2.5.1 Trip-based mobility models

One of the traditional modelling frameworks in the transport sector is the classic four-stage trip based transport mobility, where the unit of measurement of travel phenomena is trips. In the traditional four-stage model, mobility phenomenon is conceived as made up of four steps of trip generation, trip distribution, mode choice and route choice as illustrated in Figure 2.1. In practice a further equilibrium of the network of flows is sought to optimize a particular utility of the network passengers (Patriksson, 2015, Willumsen, 1981). Behavioural models and analytical methodologies are then developed for each stage to ascertain the influence of associated exogenous and endogenous passenger and network attributes.

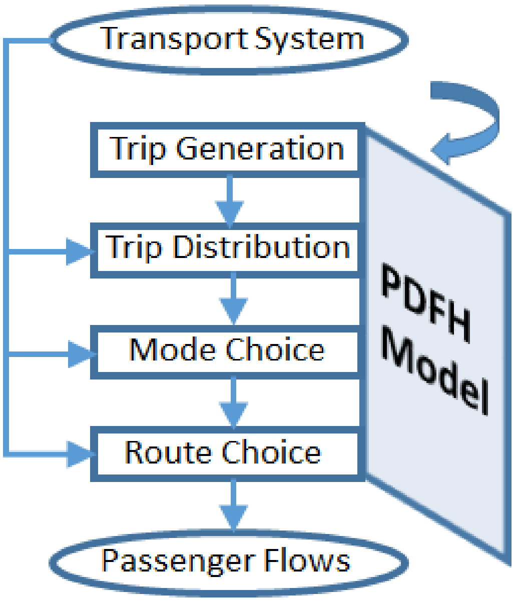


Figure 2.1 Traditional 4-stage transportation modelling concept.

2.5.1.1 Trip based mobility

Trip generation is traditionally modelled by expressions relating the socio-demographic characteristics of zones, to an ability to attract and produce flows of passengers. Measures of estimates for the zones are then normalised to ensure outward and inward flows for each zone are equal. The trip distribution follows on from the generation and cross-tabulates the number of trips originating and destined for the zones in a matrix. Various strategies have evolved over the years for modelling trip distribution, and include the gravity model (Zipf, 1946), radiation model (Simini et al., 2012), entropy maximisations models (Wilson, 1969), and a host of other so called spatial interaction models (Fotheringham, 1983, Wilson, 1971). Mode choice models then yield the choice behaviour of the population toward a particular mode of transport. Early requirements meant that diversion curves (de Dios Ortuzar and Willumsen, 1994) were adopted for splitting travel into various modes, positing that travel in monetary and time costs are the driving influences on mode choice. Disaggregate choice models have recently been put forward as an effective modelling strategy for mode choice (Ben-Akiva and Lerman, 1985a, McFadden, 1973). Route choice models (Ben-Akiva et al., 2004) appraise passengers' perception of route characteristics to forecast paths of travel likely to be chosen. Typically a multinomial logit regression is used to model the choice set of routes, with characteristics including length, time, cost, scenic nature etc., yielding an optimal choice of route.

Much of the modelling effort in the literature has concentrated on developing the individual stages of the 4-stage model. The modelling framework developed in this thesis combines traditional spatial microsimulation from geography, with transport GIS network analysis, optimized by utilizing information from a general transit feed specification (GTFS[8]), to provide an alternative strategy to the 4-stage model. The NRTS, Census and LENNON ticketing data are combined to simulate a representative population of railway passengers embedded in the wider population, removing the need to generate, distribute, and choose modes and routes for passengers. The simulated population inherits the attributes of the NRTS, Census and LENNON, leaving only an analysis stage to identify drivers of mobility.

2.5.1.2 Direct demand forecasting

While increasingly sophisticated mobility models with increased behavioural content and improved data were being developed, the railways industry sought to improve mobility models by resorting to simplified transport demand models. One such effort yielded the Passenger Demand Forecasting Handbook (PDFH), a reference handbook for railway demand forecasting in the UK, and a comprehensive guide to the state-of-practise. The PDFH effectively integrates the various stages in the traditional 4-stage model into a single model, creating a so called direct demand model as illustrated in Figure 2.1. The PDFH utilizes econometric elasticities (Worsley, 2012) to forecast the change in passenger demand due to a commensurate

[8] The GTFS enables the identification of the accurate location of the trains, and when integrated with a GIS network dataset enables the identification of a utility maximizing route of passenger travel, constrained by the service provisions of the rail network

change in endogenous and exogenous drivers of mobility (Wardman et al., 2007, Worsley, 2012). The methods developed in this research are geared toward the category of increasingly sophisticated and richer behavioural models, and are complimentary to the established PDFH by better explaining the dynamics of mobility in areas where the PDFH might have weaknesses. As a simplified elasticity based demand model (Kumar, 1980, Willumsen, 1993) the PDFH has been chosen in the rail industry in a trade-off between modelling cost, scale and complexity. As reported in literature (Chapin, 1974, Epstein and Axtell, 1996, Ettema and Timmermans, 1997, Jones et al., 1983), newer and alternative demand forecasting strategies have resulted from a need to test hypotheses relating individual behaviour to macroscopic mobility regularities, to deliver an efficient passenger service on better operated infrastructures.

The trip is a derived demand resulting from a desire to fulfil a series of activities in a specific order within a constrained space and time, but it is the trip that is adopted as the basic unit of research in many mobility studies (Zhang et al.), with less attention devoted to activity patterns and tour models. This is mainly due to the difficulties in defining activity-pattern choice sets and in coping with the size of the range of possible tour models (Bifulco et al., 2010, Bowman and Ben-Akiva, 2001), and in reconciling analytical models of mobility interaction with physical representations of the urban environment and transport infrastructure. Ultimately, it is the desire to fulfil an activity that generates the mobility, so in developing richer mobility models, activity-patterns and not trip-counts form a better unit of research (Axhausen and Gärling, 1992, Bhat and Koppelman, 1999, Bowman and Ben-Akiva, 2001, McNally and Rindt, 2007).

2.5.2 Activity-based mobility models

Activity-based discrete choice strategies (Train et al., 1987) model mobility behaviour as a series of interrelated choices. The illustration developed in this research and shown in Figure 2.2 is a disaggregate conceptual model of mobility, based on individual choice behaviour, enabling time-space geographic constraints (shown by the constrained lines of flows) to be incorporated in modelling tours, and enabling subsumed individual activity patterns (concentric circles) to be modelled. The new conceptual mobility model is individual-level and disaggregated, and mobility is conceived as broken down into stages whereby a household activity pattern for instance would subsume an individuals' activity pattern.

Within this conception, tour models are conceived so that a tour time-of-day is constrained by tour destination-choice, in turn constrained by tour mode-choice and route-choice Figure 2.2. This conception enables the incorporation of geography space-time constraints and behavioural analytical methodologies based on discrete choice models (DCM), and developed to model the mobility (Hess and Daly, 2014). DCM's are amenable for use with a wide range of data types as input, with a great flexibility in model specification, rendering them suitable to define a wealth of models (Ben-Akiva and Lerman, 1985b, Brownstone, 2001, McFadden, 1980, McFadden and Train, 2000).

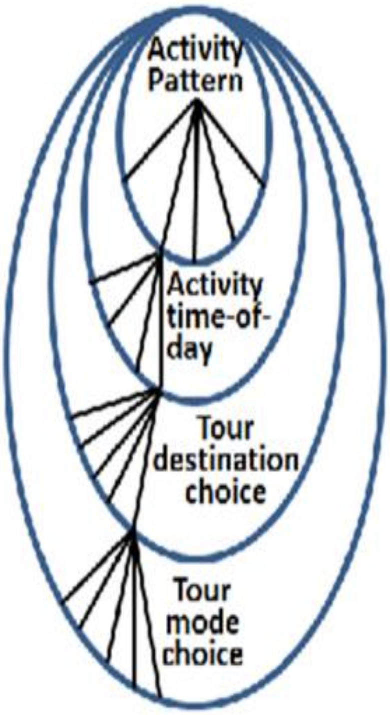


Figure 2.2 Activity-based mobility modelling concept.

In this thesis the activity modelling construct shown in Figure 2.2 is not developed any further due to the unavailability of activity schedule data to the project. However, the micro-level synthetic population created in this research contains a level of detail synonymous to that expected of modern multi-agent activity-based models. The novel developments in the spatial microsimulation and the GIS-GTFS network models proposed in this thesis facilitates accurate assignment of passengers to the transport traffic, thereby precluding the need for utility maximizing traffic assignment in traditional multi-agent models (Horni et al., 2016).

2.5.2.1 Review of multi-agent transport models

Many traditional surveys generate stated preference data (George et al., 2014), as opposed to consumer data which tends to be revealed preference. The revealed circumstantial detail in consumer data potentially enables the creation of rich contextual detail, in turn enabling causal inference on drivers of passenger demand (Grimmer, 2015, Varian, 2014). The context of big data is typically identified by replicating the physical and logistical circumstance of data creation. It is difficult and tedious to replicate a real physical system (Assad, 1980), but in the case of the rail sector, the recent availability of both pre-scheduled and real-time transit train movement information, the so called GTFS information (ATOC, 2017), enables the accurate reproduction of transit mobility and associated circumstances surrounding the passenger’s experience within the logistical GIS railway infrastructure. This forms the GIS-GTFS network model.

The multi-agent transport modelling strategies based on the traditional 4-stage model (Ortuzar and Willumsen, 2002), and implemented in software like MATSim (Gao et al., 2010, Horni et al., 2016) and EMME (Consultants, 1999), broadly consist of the following stages: demand and activity generation, modal and route choice, network and traffic simulation, and then adaptation, feedback, and learning for traffic assignment. MATSim and EMME are in the main traffic assignment tools typically targeted at the last two stages (of the 4-stages), while adopting external solutions for the first two stages. For the rail sector, an application which exploits novel consumer data for use in the spatial microsimulation process, and then identifies the context of mobility using a GIS and the GTFS information, would address the first three stages of multi-agent modelling. The final stage of traffic assignment arose in multi-agent genres because typically the simulated input demand is an imprecise reflection of passenger mobility, requiring utility equilibrium models in the assignment of flows to the transport network. With novel consumer data which when improved potentially reveals accurate and precise contextual detail of mobility, the necessity of complex utility optimizing traffic assignment may need a review. In essence, if the simulated data created is rich enough (containing information on passenger arrival time at train station, first train, intermediate stops, final destination, ticket restrictions, etc.), it would preclude the need for the added complexity of optimizing the passengers assignment to the network traffic. Creating such representative micro-level synthetic demand population of railway passenger’s forms a major part of the research in this thesis.

MATSim has not been used in conjunction with population synthesis methods that are based on large skewed consumer mobility interaction data. There are no papers in the literature that have used mobility interaction data as a basis for creating a synthetic population. MATSim being a traffic assignment model would remodel the mobility of the passengers derived from a spatial microsimulation as part of the process of creating a marginal utility transport assignment model (Horni et al., 2016). As a result data output from MATSim and such traffic assignment models strictly can no longer be seen as portraying native mobility data that can be used for say subsequent Bayesian regression and causal inference on drivers of mobility. In such a scenario, the causal inference would identify the attributes of the marginal utility model developed as part of MATSim implementation. The approach presented in this thesis excludes the complexities of remodelling the input demand data (Nagel and Flötteröd, 2012, Patriksson, 2015). However, the approach enhances such data by incorporating endogenous attributes from GIS-GTFS network simulation, in readiness for subsequent spatial analysis (by Bayesian modelling as developed in this research) (Gelman et al., 2014b, Lunn et al., 2000). The framework for harnessing typically skewed consumer data, and for including mobility interaction through a 3-level simultaneous constraint hierarchy within a spatial microsimulation, and then simulating rich exogenous and endogenous attributes using a GIS-GTFS model, has the potential to extend or alter the traditional multi-agent transport simulation genres. Applying this novel strategy to big data would be the first time such method is presented in the literature.

2.5.2.2 Alternative disaggregate models

Apart from the trip-based, activity-based and multi-agent transport modelling genres, an alternative framework is agent-based modelling. Agent based models (ABMs) are maturing as a concept for providing insights into urban phenomena as influenced by environmental and socio-demographic characteristics, with applications in transport modelling (Huynh et al., 2014, Miller et al., 2004, Raney et al., 2003, Salvini and Miller, 2005). In ABMs, insights gained from traditional transport modelling exercises are used to assign behaviour to digital agents. These agents are simulated to engage in mobility, while the pattern of mobility that evolves is studied. Traditionally ABM implementations are not aimed at replicating reality (Malleson and Birkin, 2012), but instead seeks to disaggregate the complexities of mobility into constituent parts which when combined create a richer picture of evolution of the mobility process (Heppenstall et al., 2011).

In transport ABMs, each individual or entity is modelled as an agent with behaviour, autonomy, activity schedules and household associations to simulate the evolution of mobility. ABMs are disaggregating models. The ABM transport demand model for Sydney (Huynh et al., 2014) incorporating the TRANSIMS package consists of a population synthesizer and a route planner, combined to be akin to the spatial microsimulation implemented in this thesis. The traffic micro-simulator of the TRANSIMS package is akin to the GIS-GTFS network simulation implemented in this thesis. The ABM implementations tend to include a logit choice model to objectively describe the utility maximizing decision making processes of agents. The difference in the framework for urban mobility presented in this thesis and ABM is that whilst the former aims to replicate reality, the latter does not aim to replicate reality, but rather to explore the emergent logit choice phenomena yielded by different scenarios of agents interacting. Replicating reality yields more revealing results, hence this research is a progression from ABM.

2.6 GIS logistical network models

Geographic Information Systems (GIS) are established in the literature (Goodchild, 1992, Longley, 2005), and its strength lies in the ability of GIS to scale across any physical expanse, providing a holistic view of objects in their real life context. This strength gives GIS enormous potential for use in building a model replicating a logistical rail network. The strengths of GIS technology include structured layered features, accurate feature classes, precise attribute data and topologically rich analytical models. As such the role of GIS (as in a realistic rail network model) is that of an integrator, analyser and visual thematic display (Egenhofer and Kuhn, 1999, Odiari, 2011) which seamlessly synergises the transit schedule information embedded in the GTFS (ATOC, 2017).

The reason GIS is inherently ideal as an integrator is firstly because it has the ability to scale across any expanse (geospatial, layer, schematic and temporal) ideally suited for identifying global and local trends and secondly, GIS supports a methodical homogeneous collection of common features with rich topologic relationships across features enabling a seamless application across workflows. Thirdly in GIS, annotations and descriptive attribute data are very specifically defined such that for example, if an attribute in a feature class is of a specific data type, the field will only accept accurate and precisely formatted data as inputs, resulting in strong data typing, ideally suited for precise and accurate spatial analysis (Odiari, 2011).

The GTFS information consists of detailed transit schedules, and when used in collaboration with a GIS enables a query of the precise mobility of each transit train in space-time on the rail network. The context of mobility of each passenger input to the rail network is identified by querying the passengers network routes solved using the GIS-GTFS model. Model Builders (ESRI, 2016) are used in this research to automate and simplify the GIS analysis steps. Once the model of the rail network is built, the inherent nature of GIS[9] enables an assessment of 'what-if' scenarios, enabling interventions to be assessed. As an integrator, analyser and visual thematic display (Egenhofer and Kuhn, 1999), a GIS can be extended for application as a strategic, tactical and operational railway model. Once the rich attribute micro-level representative data has been created using the first two methodologies of spatial microsimulation and GIS-GTFS network modelling, the outstanding issue becomes the missing behavioural attributes of the rail passengers. As highlighted earlier, these missing values include journey factors, ticket validity periods and flows to group stations, and are deduced by creating Bayesian models of the mobility phenomena, whereby any missing values are identified just like the parameters of the mobility model. Bayesian modelling is reviewed in the next section in the context of its application to the imputation of missing ticketing values and for use in mobility analysis, highlighting the particular attraction of such a framework.

2.7 Review of Bayesian modelling framework

Big data tend to reveal the customer choice behaviour, which can often evolve and interact with changes in the service provision. As a result caution has to be exercised in adopting traditional black box machine learning, neural networks and artificial intelligence analysis solutions that assume that a sufficient volume of data was sufficient to guarantee object inference. This perception does not take cognisance and does not incorporate knowledge of changes in service provisions which are the circumstances of the data generation, and which may present a time variant dataset. A neglect of the mechanism of data creation has been reported in the literature, leading to the so called 'big data hubris' (Bollen et al., 2011, Butler, 2013, Lazer et al., 2014). These portend to a need to understand the mechanism of the data generation process, especially in relation to the complexities of mobility, prior to data imputation or adopting an analysis methodology.

[9] Strengths highlighted above make GIS suited for interventions analysis.

2.7.1 Data imputation strategies

The choice of imputation method for each dataset depends on the statistical mechanism of omissions (Graham, 2003, Little and Rubin, 2014), data type and relationship between key variables (Honaker and King, 2010), and percentage/distribution of omissions (Allison, 2001, Dong and Peng, 2013). The concept of missing at random (MAR) is however central to current state-of-the-art imputation methods (Little and Rubin, 2002), as it enables inferences on the population to be made from incompletely observed data. A number of data imputation methods (also called infilling) have been developed over the years as widely reported in literature (Allison, 2001, Graham, 2009, Heckman, 1979, Rubin, 1976, Rubin, 1996, Schafer and Graham, 2002, Stekhoven and Bühlmann, 2012). These are broadly classified as ad-hoc, heuristic, or parametric methods. Ad-hoc methods replace the missing values directly based on an assumption about knowledge of the missing values. Heuristic and parametric imputations are principled methods as they do not assume knowledge of the missing values (by for example, assuming that they are equal to the mean of other values). Instead principled methods take the context of the observed data as the basis for estimating the missing values. Heuristic methods are non-parametric with no inherent assumptions about the distribution or model of the data, but exploit practical evolutionary learning algorithms to predict the missing data values. Parametric methods on the other hand assume a prior parametric model and distribution for the data. A summary of a range of established ad-hoc, heuristic, and parametric imputation strategies are detailed in the literature (Horton and Lipsitz, 2001, Rubin, 1996).

Ad-hoc imputation methods are attractive through being simple to implement, but involve strong unestablished assumptions about the data, like assuming the missing data are the mean of the observed data in the means method (Schafer, 1997, Schafer and Graham, 2002), or the list-wise deletion (King et al., 1998) which assumes the data are MCAR. Current industry rail ticket infilling methods adopt a logical-rules based ad-hoc method which assumes the passenger always adopts the cheapest route of travel (SDG, 2011, SDG, 2016, Taylor, 2013b). The disadvantages of ad-hoc imputation strategies in over-simplifying the imputation problem are widely reported in literature (Allison, 2001, Little and Rubin, 2002, Little and Rubin, 2014, Rubin and Schenker, 1991, Schafer and Graham, 2002, Van Buuren et al., 1999, van der Heijden et al., 2006). In the rail sector, it is important how missing data values are dealt with, as they relate to ticket-revenue allocation to TOCs

Non-parametric heuristic imputations aim to capture the complexities and non-linearity’s found in realistic mobility data, by employing machine learning strategies (Shah et al., 2014). A change in train access or time-table resulting in a change in cross-sectional demand parameters in time can be better captured by a model that allows time-trend shifts across period cross-sections of data. True mobility phenomena derived from contextually rich datasets include complexities and non-linearity’s (Giannotti et al., 2011) not easily captured by default parametric and semi-parametric models, resulting in a need for suited machine learning strategies. A range of heuristic imputation methods are established in literature (Batista and Monard, 2002, Siddiqui and Ali, 1998, Stekhoven and Bühlmann, 2011). A disadvantage of heuristic methods lies in their complexity and inability to replicate the data generation process and models. The predictive mean matching method (Durrant, 2005) often used in heuristic strategies, for instance infills randomly from observed values, such that in a hierarchical dataset with observations missing in one of the hierarchies, the imputation will consistently fail.

Parametric methods typically involve creating a regression model representing the mechanism of data generation, and using this as a basis for imputing missing values (Enders and Bandalos, 2001, Gelman et al., 2014b, Rubin and Schenker, 1991, Van Buuren, 2007). The parametric methods differ by including models in the imputation process to reflect the mobility phenomena. The MICE (Van Buuren, 2007), EMB (Honaker et al., 2011) and BUGS (Spiegelhalter et al., 1996) are attractive imputation strategies that enable the uncertainty associated with the imputation to be estimated. The weakness in MICE methods are the implicit MAR assumption that are not ascertained or substantiated. The EMB and BUGS are suited to realistic problems with a large number of co-variates (typically >20), while BUGS further enables easy creation of models better reflecting reality (Lunn et al., 2012, Plummer, 2003), making the Bayesian framework the choice strategy.

2.7.2 Bayesian analysis strategies

Research has shown that the particular mobility model adopted (whether variants of levy flight, random walk or the more traditional models (Gonzalez et al., 2008, Simini et al., 2012), heavily impinges on the accuracy of the data upon which the forecasts are based. Imputing missing values is conceptually equivalent to forecasting unknown values, and as such the need for inclusion of an appropriate mobility model for forecasting cannot be overstated for imputation applications. The particular model adopted has to take cognizance of the data generation process, especially in consumer data applications where data are not collected purposefully but revealed from consumer responses to typically evolving service provision. As mentioned earlier, this was poignantly brought to the fore in the ‘big data hubris’, where it was assumed that being large (50 million records) was self-sufficient to provide correct inferences without taking account of the data mechanism and generation process, in reconciling with a stated survey of about 1,152 data points (Butler, 2013, Lazer et al., 2014).

Bayesian modelling is increasingly gaining prominence for data imputation and analysis, as many of the advance ad-hoc and heuristic methods increasingly adopt Bayesian procedures to solve their intractable expressions (Roy et al., 2017, Paddock, 2002). Bayesian strategies have potential for application to the mobility on the railways which tend to be non-linear and complex (Krajzewicz et al., 2002). To put these into perspective with an example, rail passengers are attracted to season tickets because they can be used multiple times daily, also tending to be cheaper commuter fares (Hensher and Bullock, 1979). One-way and single return tickets on the other hand have truncated uses restricted to one single and one return journey respectively. All passengers are considered identically distributed (Clauset, 2011) by coming from the same pool (the distribution of railway passengers). However, the passengers clustered by ticket product can also be described as exchangeable in the statistical sense (De Finetti, 1972), since the behaviours across the groups are independent subject to belonging to the same distribution of railway passengers. As a result of these complexities, mobility on the railways would be better modelled as hierarchical mixtures, best represented within the Bayesian framework (Gelman et al., 2014b), making Bayesian modelling the particularly choice strategy (McElreath, 2015).

The rate of use of season tickets would depend on the length of journeys the passenger makes. Due to the limited 24 hours in the day, passengers on lengthy journeys can only repeat these fewer times than passengers making shorter trips. An adequate imputation method would need to reflect the potential differences in daily journey rates (journey factors) associated with the distances the passengers travel. It is further known that daily trip rates of passenger also relates to the age bracket of the passenger (DfT, 2013b). Bayesian framework has simple constructs that enable the censoring of journey factors subject to the passengers travel distance (Lunn et al., 2012, Plummer, 2003). Another set of tickets with missing values are those with flexible periods of validity. For instance, the ticket product *MTL* is valid for between 90-180 days, and whilst a fare is paid, no information about the specific number of days of validity of these tickets is recorded. As such in imputing these values, an adequate model would need to apply a 90-180day truncated limit on the imputed values for the ticket validity. The Bayesian framework has such simple constructs for truncating the validity periods of flexible tickets. This facilitates the inclusion of such industry knowledge garnered from a wealth of empirical railway experience.

Within the Bayesian modelling framework the strength of the analyst is an ability to easily and flexibly specify an adequate representative model that captures the relevant facets of mobility. Bayesian models are particularly suited to hierarchical phenomena (Raudenbush and Bryk, 2002) associated with the time-series cross-sectional nature of the ticketing data. Within this framework, the Bayesian methodology enables data augmentation and imputation by weights (Horton and Kleinman, 2007), suitable for application to the imputation of flows to grouped stations (and to PTE ticket use). Within the Bayesian framework, all the constructs can be brought to effective use with a few simple steps, implemented within a mobility model, thereby taking cognizance of the mechanism that generated the data in the first instance. Based on the above, there is a compelling reason to adopt Bayesian modelling for imputation of railway ticketing data and for mobility analysis.

2.8 Chapter summary

The ATOC railway ticketing data are available within the LENNON ticketing database, representing tickets sold on the UK rail sector. Integrating these with other relevant datasets and feeding through a logistical rail network yields a dataset rich in endogenous and exogenous attributes, a pre-requisite for objective imputation (Rubin, 1976). A Bayesian imputation identifies the missing idiosyncratic behaviour of passengers, thereby creating an improved dataset useful for subsequent mobility analysis (Viktor and Kenneth, 2013). The concepts applied in the project and reviewed in this chapter include three methodologies developed to harness railway ticketing consumer data for utilization in mobility analysis. The project stages are summarised in the flowchart of Figure 2.3.

In the next chapter, the setting and a brief description of the study area are presented, as well as the data pre-processing. The strategy is presented for identifying the nature of other relevant datasets to combine with consumer ticketing data. The preliminary data pre-processing stages are presented, as well as a preliminary summary of the ticketing data.

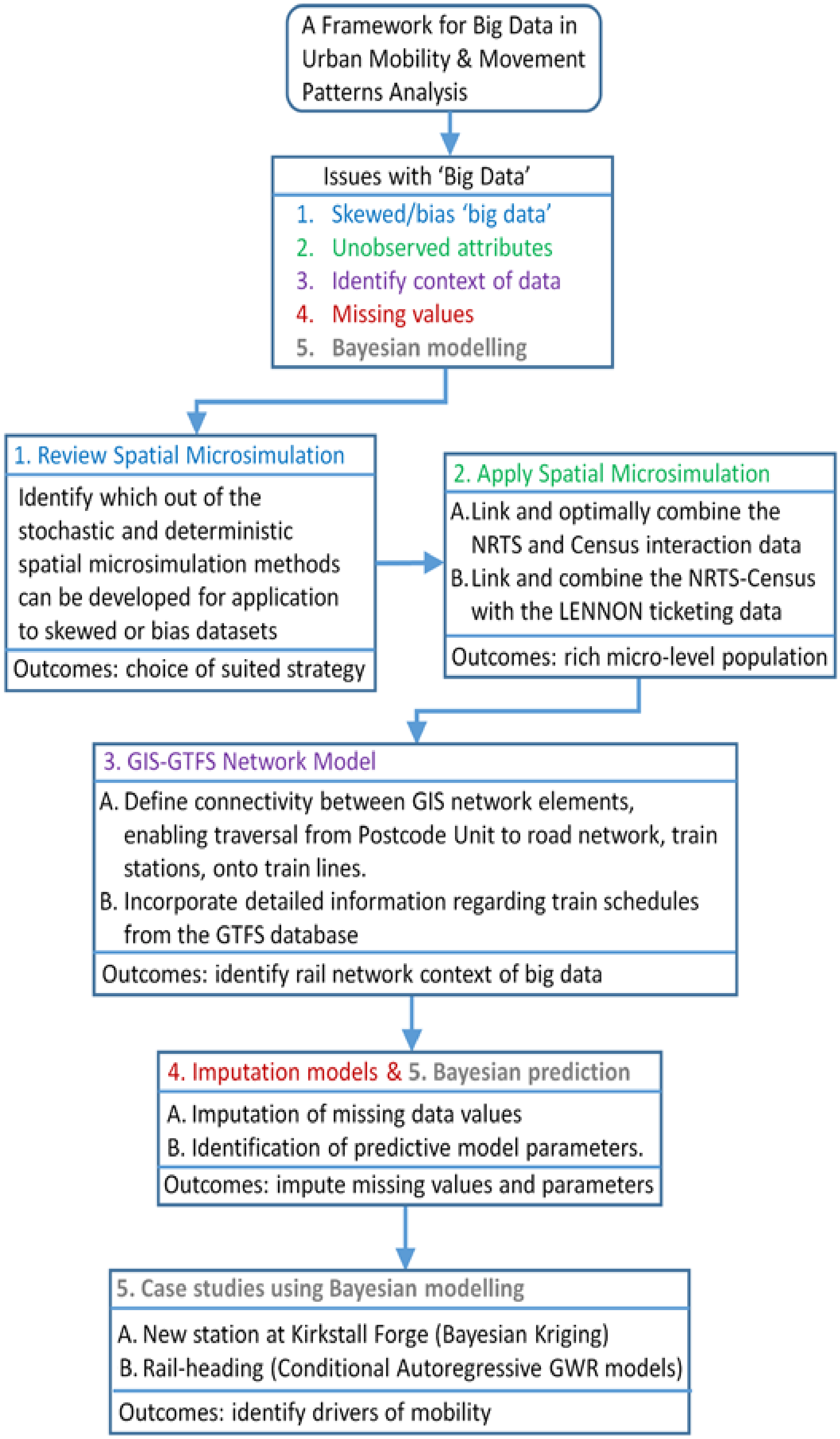

A Framework for Big Data in Urban Mobility & Movement Patterns Analysis
Issues with 'Big Data'
1. Skewed/bias 'big data'
2. Unobserved attributes
3. Identify context of data
4. Missing values
5. Bayesian modelling
1. Review Spatial Microsimulation
Identify which out of the stochastic and deterministic spatial microsimulation methods can be developed for application to skewed or bias datasets
Outcomes: choice of suited strategy
2. Apply Spatial Microsimulation
A. Link and optimally combine the NRTS and Census interaction data
B. Link and combine the NRTS-Census with the LENNON ticketing data
Outcomes: rich micro-level population
3. GIS-GTFS Network Model
A. Define connectivity between GIS network elements, enabling traversal from Postcode Unit to road network, train stations, onto train lines.
B. Incorporate detailed information regarding train schedules from the GTFS database
Outcomes: identify rail network context of big data
4. Imputation models & 5. Bayesian prediction
A. Imputation of missing data values
B. Identification of predictive model parameters.
Outcomes: impute missing values and parameters
5. Case studies using Bayesian modelling
A. New station at Kirkstall Forge (Bayesian Kriging)
B. Rail-heading (Conditional Autoregressive GWR models)
Outcomes: identify drivers of mobility

Figure 2.3 Complementary technologies used in project.

Chapter 3
SETTINGS, DATA AND PRE-PROCESSING

This chapter describes the geographic area of the county of West Yorkshire from which the datasets are derived, and briefly overviews the existing transport infrastructure. This chapter contextualizes the consumer ticketing data used for mobility analysis on the railways. The volume and nature (velocity, variety, and veracity) of the consumer ticketing data aptly defines it as ‘big data’, making its use appropriate for developing and validating a framework for the use of such datasets in urban mobility and movement pattern analysis. The terminology used to describe the statistical missing data mechanism in the rail sector dataset is presented in more detail. The detail of pre-processing stages required to unravel the data are presented, as these facilitate a better understanding of the mechanism through which the missing values occurred. The criteria that enable an assertion of the necessary covariates required for objective imputation is developed. Little’s test for missing values is presented as that strategy for identifying the relevant covariates. Little’s test indicates the variables which when added to the database would explain the difference between the observed and missing data values, thereby making the missing values MAR.

The framework presented in this thesis harnesses the railway ticketing data by optimal combination with other relevant datasets. To facilitate combining of the datasets used in the research, the geographies of the datasets need to be aligned and reconciled. The re-zonation method adopted for these are presented (Openshaw and Rao, 1995). To facilitate the process of relating the datasets, the covariate of these data need to be associated. When no direct association exists, a relation can be achieved in many instances by creating a new variable by restructuring an existing one (or by a combination of others). To illustrate this, the deduction of a useful income variable for the Census using an occupation variable is presented. Then, a preliminary summary of the datasets are presented. A 3-D origin-destination plot gives an indication of the flow volumes associated with each train station, as well as the potential effect of an imputation on these volumes. A classification of the continuous and categorical variables and values is used to preview relationships between disparate data variables and values. Further a spatial interaction regression model on the complete case dataset is used to give a preliminary insight into the mobility phenomena.

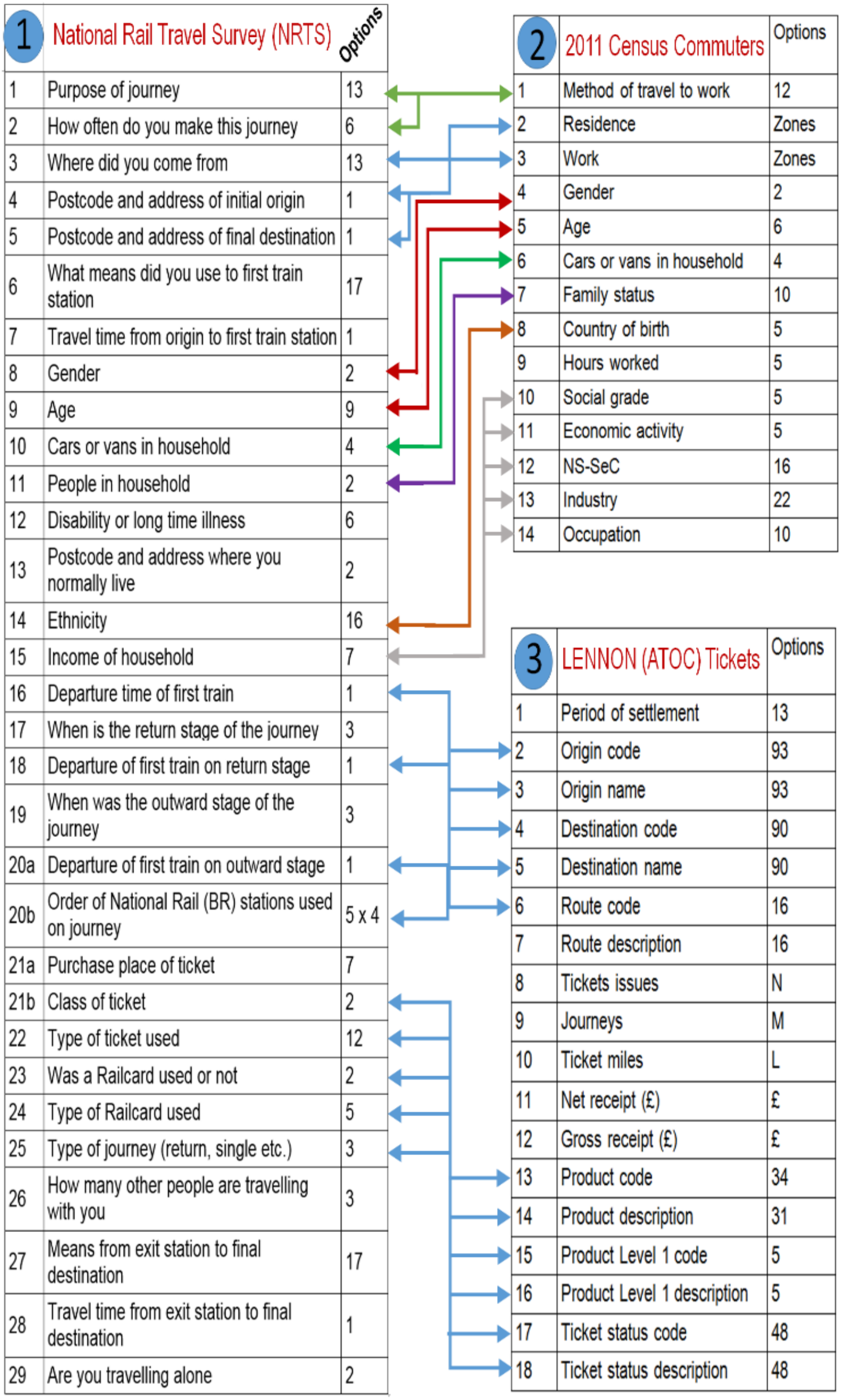
3.1 West Yorkshire study data

The analysis in this thesis is focused on West Yorkshire (see Figure 1.1), an inland metropolitan county made up of five districts in the north central of UK, with a population of 2.2 million, and covering 2030km^2. The largest hub is the city of Leeds, which includes Leeds Bradford International Airport, major railways and three major motorways that traverse the county. Leeds is also a planned location for a terminus or a major spur hub for a proposed high speed railway development in the UK. The East Coast Main Line is a key rail artery on the eastern side of Great Britain which connects into the Channel Tunnel and Europe through Leeds and London. The West Yorkshire Integrated Transport Authority (WYCA-ITA) is responsible for developing the strategy for the existing integrated transport system in the county serving about 20million rail passengers monthly. The methodology developed in this thesis is scalable and generalizable to other rail transport networks in areas within the UK, or in other countries where large consumer revealed datasets and similar background stated choice surveys are available. The datasets combined were the 2011 Census interaction data (ONS, 2013), the National Rail Travel Survey (NRTS) (DfT, 2013a) and 'big data' consisting of railway ticket sold in the West Yorkshire study area. The ticketing data were procured from the Association of Train Operating Companies' (ATOCs) LENNON ticketing database (ATOC, 2003b). The relational table illustrating the association of the attributes in each dataset is shown in Figure 3.1.

The LENNON dataset is described as time-series cross-sectional (TS-CS), whereby a tab is kept on each ticket sale within each 4 week period forming a cross-section. Thirteen (13) sequential four-week time periods are defined over each year and the data are partitioned according to the time period of procurement of individual tickets. Within each time period, the dataset is assumed independent and identically distributed (***iid***). The independence of the data means that each ticket sale is independent of all others. The identical distribution means that each data point is sourced from the pool or distribution of UK railway passengers. In essence, if a passenger buys a train ticket, it does not affect or influence the ability of another passenger to buy a train ticket, seemingly assuming that the tickets are procured from an infinitely large pool of tickets or picked with replacement. By definition an ***iid*** dataset is exchangeable (but not necessarily vice versa), and as such ticketing data within a period are statistically exchangeable (De Finetti, 1972, Kingman, 1978, Koch et al., 1982).

On this basis, the LENNON ticketing data are re-classified into cross-sectional sets over each 4 week time period, enabling the analysis of each cross-section. This rests on the big assumption that ticket purchasing patterns across each period lasting 4 weeks (approximately a month) are stable without fluctuations within the period. Any seasonal changes are assumed to occur across seasonal periods, i.e. across the months with seasonal fluctuations on a monthly basis. On the strength of this assumption, the LENNON ticketing data are TS-CS, and within each period (CS) the data are statistically exchangeable, presuming that the mobility phenomena within each period are fundamentally similar (De Finetti, 1972, Gelman et al., 2014a), with nominal changes due to variations in service provision, seasonal variation, and the like. In this thesis, due to expedience, a cross-section of the ticketing data are analyzed, although the time-series feature of large consumer data are an attractive feature for analysis to identify trends in mobility phenomena. In a wider context, the time-series (TS) feature of novel big-data can be accommodated within an exchangeability modelling framework by considering each time-series set as a hierarchy, whereby the mobility phenomena is modelled as a mixture of exchangeable cross-sections. In the literature, a range of models have been developed to accommodate TS-CS datasets within the Bayesian framework based on autoregressive and non-stationary data models (Beck, 2008, Beck and Katz, 1995, Brunsdon et al., 1996, Fotheringham et al., 2003, Lunn et al., 2012).

Another relevant rail sector dataset used in the research are the General Transit Feed Specification (GTFS) which are detailed transit schedule information sourced from ATOC (ATOC, 2017) (but typically also available from transport agencies (TransitFeeds, 2017)). The GTFS describes network trips and their geographic shapes, routes, stops, and stop-times, schedule calendar dates and exceptions, transfers, service frequencies, and feed metadata (Google, 2016). This information facilitates the detailed specification of the transit movement (scheduled and in real-time).The variables within the data tables of the GTFS information are represented by the Class Diagram shown in Figure 3.2, with further details available in the literature (Odiari, 2018 (awaiting review)). The GTFS information is published at time intervals ranging from between 5 to 9 days, and at the end of each monthly ticketing period or after each fare round (ATOC, 2017).



1	National Rail Travel Survey (NRTS)	Options
1	Purpose of journey	13
2	How often do you make this journey	6
3	Where did you come from	13
4	Postcode and address of initial origin	1
5	Postcode and address of final destination	1
6	What means did you use to first train station	17
7	Travel time from origin to first train station	1
8	Gender	2
9	Age	9
10	Cars or vans in household	4
11	People in household	2
12	Disability or long time illness	6
13	Postcode and address where you normally live	2
14	Ethnicity	16
15	Income of household	7
16	Departure time of first train	1
17	When is the return stage of the journey	3
18	Departure of first train on return stage	1
19	When was the outward stage of the journey	3
20a	Departure of first train on outward stage	1
20b	Order of National Rail (BR) stations used on journey	5 x 4
21a	Purchase place of ticket	7
21b	Class of ticket	2
22	Type of ticket used	12
23	Was a Railcard used or not	2
24	Type of Railcard used	5
25	Type of journey (return, single etc.)	3
26	How many other people are travelling with you	3
27	Means from exit station to final destination	17
28	Travel time from exit station to final destination	1
29	Are you travelling alone	2

2	2011 Census Commuters	Options
1	Method of travel to work	12
2	Residence	Zones
3	Work	Zones
4	Gender	2
5	Age	6
6	Cars or vans in household	4
7	Family status	10
8	Country of birth	5
9	Hours worked	5
10	Social grade	5
11	Economic activity	5
12	NS-SeC	16
13	Industry	22
14	Occupation	10

3	LENNON (ATOC) Tickets	Options
1	Period of settlement	13
2	Origin code	93
3	Origin name	93
4	Destination code	90
5	Destination name	90
6	Route code	16
7	Route description	16
8	Tickets issues	N
9	Journeys	M
10	Ticket miles	L
11	Net receipt (£)	£
12	Gross receipt (£)	£
13	Product code	34
14	Product description	31
15	Product Level 1 code	5
16	Product Level 1 description	5
17	Ticket status code	48
18	Ticket status description	48

Figure 3.1 NRTS, Census, and LENNON relational table

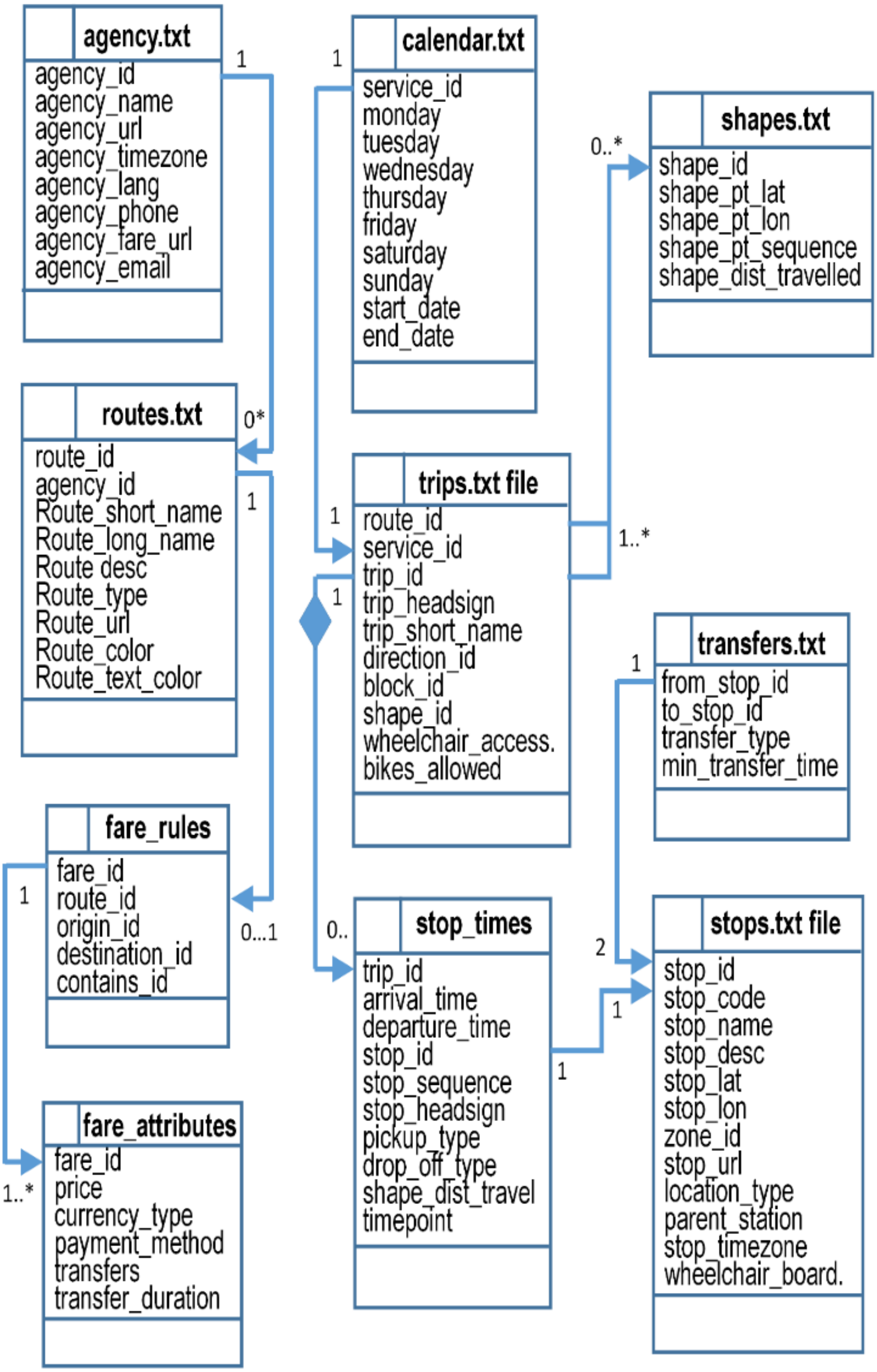


Figure 3.2 Class Diagram for the railways GTFS information showing the relational variables.

3.2 Railways statistical missing data

The formalisation of the terminology used to describe the statistical missing data mechanism in the rail sector dataset is presented. This formalisation puts the missing value concept in a context that can be related to Rubin's work (Rubin, 1976) as applicable to the rail sector ticketing data. Three points forming a corollary to Rubin's postulation, and which are a prerequisite for valid inference on datasets with incomplete values are: 1) when data are not observed at random, they are not representative of the populace being sampled, 2) the mechanism that leads to missing data has to be distinct from mobility mechanism being inferred, and 3) additional relevant variables have to be sought and linked to the observed data to explain the difference between the measured and missing values thereby making the data (MAR). The relevance of these in the context of the rail sector ticketing data are presented in the following section.

3.2.1 Missing data mechanism

Season tickets valid for travel between an entry and an exit station may typically be used several times in a day. The passenger choice of ticket usage would depend on the activity the passenger wants to fulfil which is related to their socio-economic, demographic, urban morphology, station access and egress, and other network variables. Conceptually, if prior knowledge exists of the unique passenger characteristics (i.e. parameters) $\theta,$ that are likely to influence (cause or yield) a known distribution of season ticket usage $x \epsilon X,$ then it is possible to estimate the ticket usage (journey factor, station entry, exit etc.) from the given set of parameters. A formal corollary is that the likelihood of a set of parameters θ given a set of outcomes $x,$ is by definition equal to the probability of the observed outcomes given those parameter values. This is statistically defined as $\mathcal{L}(\theta|x) = P(X = x|\theta),$ the so called likelihood function, where $\mathcal{L}$ stands for the likelihood, and P the probability. In the context of ticketing data, the outcome attributes X would be made up of an observed component X_{OBS} and a missing component X_{MIS}. In the situation where the observed values of X_{OBS} capture the entire essence of the system parameters θ, the data are missing at random (MAR). For example, if the choice of rail entry station is solely dependent on access facilities provided at stations, then any omission within the station entry variable will be MAR provided access facilities are observed. If access is not observed, then any omission in station entry variable would be classed as missing not at random (MNAR) as the ***missingness*** is dependent on unobserved (missing) station access variable.

The names MAR and MNAR are not very intuitive, but the **random** referral (as briefly described in Chapter 2) simply implies that the missing values can respectively be considered distributed **randomly** or **not randomly** with respect to the observed variables. In the example given in the previous paragraph, for passengers presented with the same access facility, the likelihood of buying a season ticket is random. If there are pertinent considerations made in buying a season ticket (like speed of journey), then if that consideration is not observed, the missing values in the other covariates would be considered MNAR (see also Section 2.3.1).

Formally, let the LENNON dataset X be partitioned into an observed part X_{OBS} and the missing part X_{MIS} , whereby $X = (X_{OBS}\ ,\ X_{MIS})$. If M is defined as the matrix of ***missingness*** with same dimension as $X,$ with elements 1 or 0 depending on whether corresponding elements in X were observed (1) or missing (0). When data are missing, the probability model to describe data is the full joint probability given by $P(X_{OBS}, X_{MIS}, M|\theta, \varphi)$, where φ is the parameter of ***missingness.*** Integrating over the missing values X_{MIS}, gives:

$$\int P(X_{OBS}, X_{MIS}, M|\theta, \varphi)\ dX_{MIS} \\ = \int P(M|X_{OBS}, X_{MIS}, \varphi)\, P(X_{OBS}, X_{MIS}|\theta) dX_{MIS} \qquad (1)$$

Observe the first term on the right (after the equal sign and under the integral). It is the distribution of M, that is $P(M|X, \varphi)$. Assuming that the data are statistically missing at random (MAR), then the expression simplifies to $P(M|X, \varphi) = P(M|(X_{OBS}, X_{MIS}), \varphi) = P(M|X_{OBS}, \varphi),$ as the ***missingness*** depends only on the observed data (Little and Rubin, 2002, Schafer, 1997). If in addition to the data being MAR, that further φ and θ are distinct (this is the case if the mechanism that led to the data not being observed is distinct from the mobility mechanism being inferred), then, with φ independent of the LENNON data parameter θ, the full joint distribution reduces to:

$$P(X_{OBS}, M|\theta, \varphi) = P(M|X_{obs}, \varphi) \int P(X_{OBS}, X_{MIS}|\theta)\, dX_{MIS} \qquad (2)$$

$$P(X_{OBS}, M|\theta, \varphi) = P(M|\varphi, X_{OBS})\, P(X_{OBS}|\theta) \qquad (3)$$

$$\therefore \qquad P(X_{OBS}|\theta) \propto \mathcal{L}(\theta|X_{OBS}) \qquad (4)$$

The profound inference from the above is that the likelihood of θ given only the observed data is equivalent to the full joint probability, and under such circumstances, inference on θ based on the observed data X_{OBS} would be valid and objective, provided the data are MAR, and that the cause of the data being missing is independent of the mechanism (of mobility) being inferred. On the railways, the unobserved journey factors for season tickets are simply because there are no sensors on the network, and the mere fact that these values are not observed does not affect a passenger's mobility or use of say a season ticket. In such circumstance φ and θ are independent.

The implication of the presentations from Equation (1) to (4) are that the observed components of the rail ticketing data are sufficient to make inference about the drivers of mobility phenomena, and by implication for use in objective imputation[10]. In practical applications, all that is required is to ensure that the data are MAR, by adding relevant data variables that explain the difference between the observed and missing values. In the next section, the unravelling of LENNON data is presented to expose the mechanism of missing data, as a precursor for utilizing Little's test to ascertain the nature of the variables that would make LENNON ticketing data MAR.

3.2.2 LENNON pre-processing

LENNON consists of discrete non-negative counts of journeys per ticket (see Table 2.1). Each tuple in the LENNON dataset is a sale, whilst the number of ticket issues for that tuple is assumed to be an in-house aggregation of the ticketing data to enhance data anonymization, more especially as the tuple reflects the same ticket product. For the purposes of analysis the ticket issues associated with each tuple is treated as a unique count (frequency) of passengers. To analyze the process leading to the travel counts, it is necessary to normalize the journeys per ticket by dividing through by the period of validity of the ticket so that each ticket can yield an equally likely *rate*[11] of use. This produces the average daily rate of ticket use (journey

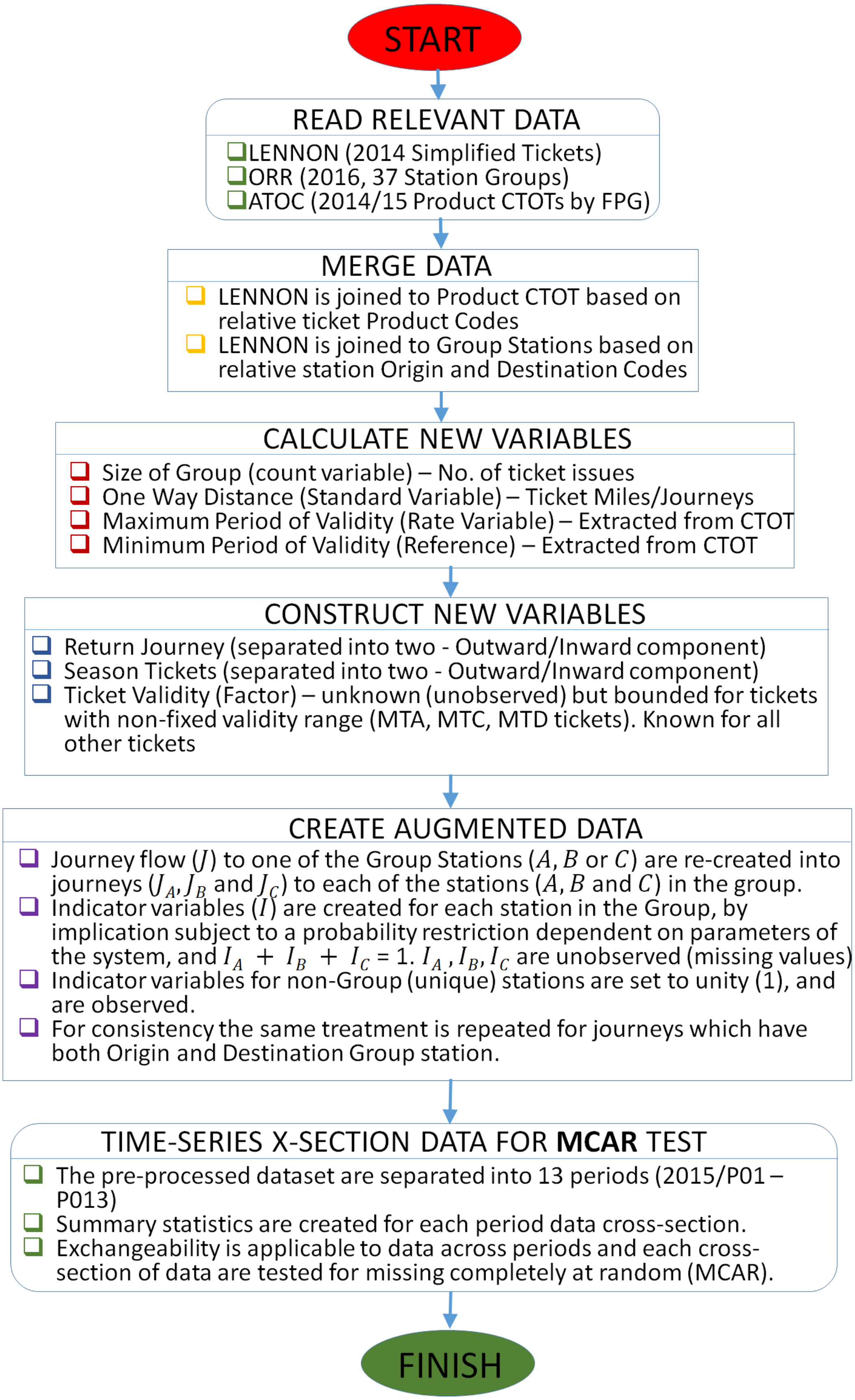
[10] As pointed out in Rubin's work, the proviso is that the mechanism that led to the missing data (i.e. the non-availability of sensors on the rail network to capture every passenger), is distinct from the mobility mechanism being inferred.

[11] Predictive mean matching (Landerman et al., 1997) algorithms which are central to many multiple imputation strategies would fail if the rate phenomena in mobility are not all individually identified and offset accordingly.

factor). Each ticket then has the same relative exposure (Dalgaard, 2008). Whilst the journey factor is a rate, an exposure (offset) variable is created equal to the period of ticket validity (Gardner et al., 1995). Certain tickets have a range of days of validity as opposed to a specific validity (for example MTA, STD and MTC/D/F product code tickets are respectively valid for between 30-60, 90-180 and 181-359 days). As such there are bounds on the period of validity of certain tickets, which fall between certain known ranges. To account for this process, the maximum ticket validity period is included as the ticket exposure, and a validity indicator is created (Landerman et al., 1997) to reflect when a ticket has a non-specific range of validity. A validity variable (a rate) is created whereby flexible tickets are considered unobserved, whilst fixed validity tickets are considered observed.

To resolve the season ticket directionality in LENNON, return journeys on season tickets from one origin to another destination (OD pair), say from Leeds to Morley have outward and inward components created in the database to account for the journey being two-way. This is distinct from current industry practice of deducing the volume of season tickets in each direction by redistributing OD return tickets based on data distributions within NRTS and PTE surveys (Taylor, 2013a). By resolving the directionality of season tickets, the basic mobility processes are effectively captured and reflected in the dataset by treating each leg of a return journey distinctively. (An additional ID information is created to identify the association of these return flows such that the inward part of the journey can only happen sequel to the outward part. Further, the description of the ticket product, for example say a day return, would impose a restriction that both legs of the ticket occur on the same day). Finally, the unit journey cost is derived by normalizing the gross revenue receipts (dividing by ticket issues and journeys made). The resulting journey cost is treated as an interaction variable, just like the distance associated with each journey.

Tickets sold to grouped stations can have flows to any one of the stations within the group. To account for this phenomena augmented flows are created for each possible origin station (and similarly for each possible destination station) within a group station cluster, with weights which would subsequently be imputed subject to a probability constraint that only one origin (and one destination) station would finally be chosen by the passenger (Horton and Laird, 1999). The sum of each set of augmented weights would add to one. Figure 3.3 details the pre-processing steps applied to LENNON to better reflect the steps in the passenger mobility phenomena.



START

READ RELEVANT DATA

- □ LENNON (2014 Simplified Tickets)
- □ ORR (2016, 37 Station Groups)
- □ ATOC (2014/15 Product CTOTs by FPG)

MERGE DATA

- □ LENNON is joined to Product CTOT based on relative ticket Product Codes
- □ LENNON is joined to Group Stations based on relative station Origin and Destination Codes

CALCULATE NEW VARIABLES

- □ Size of Group (count variable) – No. of ticket issues
- □ One Way Distance (Standard Variable) – Ticket Miles/Journeys
- □ Maximum Period of Validity (Rate Variable) – Extracted from CTOT
- □ Minimum Period of Validity (Reference) – Extracted from CTOT

CONSTRUCT NEW VARIABLES

- □ Return Journey (separated into two - Outward/Inward component)
- □ Season Tickets (separated into two - Outward/Inward component)
- □ Ticket Validity (Factor) – unknown (unobserved) but bounded for tickets with non-fixed validity range (MTA, MTC, MTD tickets). Known for all other tickets

CREATE AUGMENTED DATA

- □ Journey flow (J) to one of the Group Stations (A, B or C) are re-created into journeys (J_A, J_B and J_C) to each of the stations (A, B and C) in the group.
- □ Indicator variables (I) are created for each station in the Group, by implication subject to a probability restriction dependent on parameters of the system, and $I_A + I_B + I_C = 1$. I_A, I_B, I_C are unobserved (missing values)
- □ Indicator variables for non-Group (unique) stations are set to unity (1), and are observed.
- □ For consistency the same treatment is repeated for journeys which have both Origin and Destination Group station.

TIME-SERIES X-SECTION DATA FOR **MCAR** TEST

- □ The pre-processed dataset are separated into 13 periods (2015/P01 – P013)
- □ Summary statistics are created for each period data cross-section.
- □ Exchangeability is applicable to data across periods and each cross-section of data are tested for missing completely at random (MCAR).

FINISH

Figure 3.3 Flowchart for pre-processing LENNON.

PTE tickets were not available to the research project, however in concept, these PTE Metrocard and Day/Metro Rover tickets are treated similar to the tickets for flows to grouped stations. PTE tickets have a cap on the number of stations where they are valid for use like season tickets. The difference being that PTE tickets tend to have a wider choice set of associated origin/destination stations, and a wider choice of transport modes (feeder bus services as well as trains). Had PTE tickets been available, the mobility phenomena associated with PTE tickets could be accounted for by creating augmented data within the base LENNON dataset for each of the combinations of possible flows associated with each ticket, as was achieved with group station tickets.

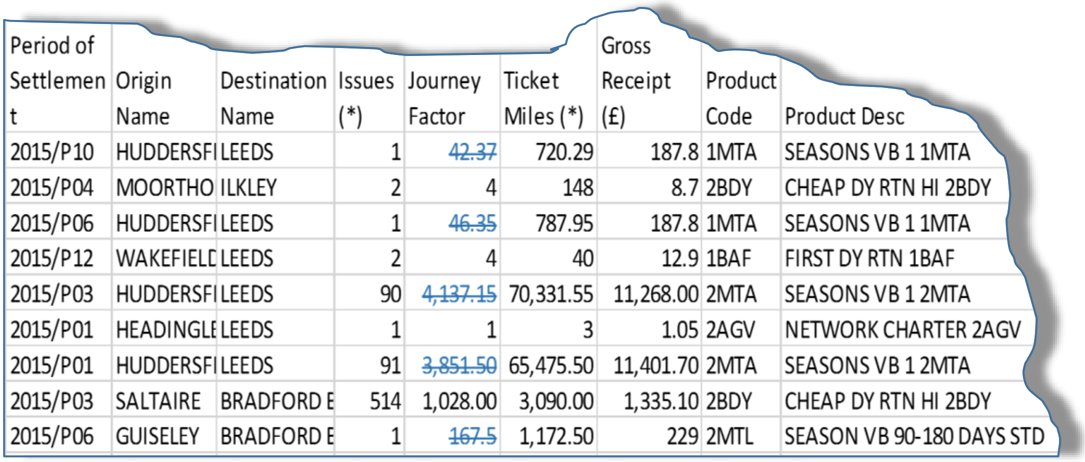
For practical applications, the more the relevant variables that are observed, the increase in the likelihood of the ***missingness*** to be ascribed as MAR (Rubin, 1976), enabling an exploitation of the MAR assumptions to proffer strategies for identifying unbiased parameters and for imputing missing values. Despite the non-existence of procedures for inferring conclusively that a dataset is MAR, there are procedures developed in the literature (Little, 1988) which have been found useful in indicating the type of variables that would complement a variable with missing values. Once this complementary relationship is inferred, it is now possible to look for additional variables that capture a similar range of phenomena as the complementary variable. This concept is hereby explained with a practical scenario:

In the LENNON dataset, the journey factor is likely to be missing (unknown or unobserved) if the ticket is a ***season*** ticket as illustrated by the struck through[12] current industry journey factor estimates in Table 3.1. A realistic estimate of season ticket daily rate of use (journey factors) ought for instance to incorporate additional effects like the length of the journey made with the season ticket or size of group associated with the ticket. It can be envisaged that much lengthier journeys would less likely be made many times in the day if the group sizes are large. Multiple daily journeys can be more readily accomplished if the journey is a short distance. Perhaps also, exogenous socio-demographics like the wealth or income of a passenger and ticket type

[12] The journey factors that have been struck through correspond to estimates using current industry ad-hoc and logical rule based strategies. The current methods adopt dated proxy trip rates for journey factor, with no consideration for evolving mobility dynamics.

(1st class or standard) would influence whether the value of the passengers time allows for frequent daily train trips. Endogenous (within-network) factors like number of changeovers made during a trip, or the ticket season of use (i.e. period of settlement which can affect the daily trip rates). The tendency is that these other covariates that influence the daily rates of ticket use would have a level of correlation with the journey factor variable. The Little's test is similar to a t-test, and assesses a measure of the relationship between the covariates in the ticket database. Sets of covariates that relate best with variables with missing values are an indication of the type of variables that would explain the unobserved missing values. Such well correlated variables are best for inclusion to make the ticketing data MAR. Little's ***missingness*** test is implemented below on covariates within LENNON.

Table 3.1 Current industry journey factor estimates.



Period of Settlement	Origin Name	Destination Name	Issues (*)	Journey Factor	Ticket Miles (*)	Gross Receipt (£)	Product Code	Product Desc
2015/P10	HUDDERSFI	LEEDS	1	~~42.37~~	720.29	187.8	1MTA	SEASONS VB 1 1MTA
2015/P04	MOORTHO	ILKLEY	2	4	148	8.7	2BDY	CHEAP DY RTN HI 2BDY
2015/P06	HUDDERSFI	LEEDS	1	~~46.35~~	787.95	187.8	1MTA	SEASONS VB 1 1MTA
2015/P12	WAKEFIELD	LEEDS	2	4	40	12.9	1BAF	FIRST DY RTN 1BAF
2015/P03	HUDDERSFI	LEEDS	90	~~4,137.15~~	70,331.55	11,268.00	2MTA	SEASONS VB 1 2MTA
2015/P01	HEADINGLE	LEEDS	1	1	3	1.05	2AGV	NETWORK CHARTER 2AGV
2015/P01	HUDDERSFI	LEEDS	91	~~3,851.50~~	65,475.50	11,401.70	2MTA	SEASONS VB 1 2MTA
2015/P03	SALTAIRE	BRADFORD E	514	1,028.00	3,090.00	1,335.10	2BDY	CHEAP DY RTN HI 2BDY
2015/P06	GUISELEY	BRADFORD E	1	~~167.5~~	1,172.50	229	2MTL	SEASON VB 90-180 DAYS STD

3.2.3 Little's MCAR test

The Little's MCAR test (Little, 1988) reveals if unique patterns exists in the missing data to assess if the probability of missing an observation is dependent on the observed and/or missing values. Significantly large p values ($p \geq 0.05$) from Little's test indicates weak evidence against the null hypothesis that the data are MCAR. For the purposes of Little's test, the journey factors in the LENNON dataset are normalised such that all journeys that have fixed observed periods of validity (single tickets or day returns) have a normalised journey factor of 1. Tickets types that have a range of validity are dealt with by assuming that they are fully used at least up until the minimum ticket-type period of validity, thereafter the remaining validity is classed as missing. However, the precise journey factor value of 1 set for tickets with fixed validity periods does not account for the uncertainty and stochasticity in real situations, as for instance some business and leisure passengers use season tickets on average more often than once a day, whilst commuters use them less often than an average of once a day over a 4-week period (HC-700-II, 2006). This unwarranted precision and no variability lead to singularity failures of the Little's MCAR test. As such a small normal random effect $(\mu = 1, \sigma = 0.05)$, in line with empirical evidence (SDG, 2011, SDG, 2014), is applied to the normalised journey factors for single and return day ticket, and similarly to factors for flexible tickets (up until the minimum period of validity).

There are missing data values in the variables of journey factor, ticket validity and group station indicators. To reveal any unique patterns in the missing data values within these variables, Little's MCAR test is applied to each pair of fully observed variable and the variables with missing values. This assesses whether the missing data mechanism is dependent on a variable within the dataset. The plots in Figure 3.4 and Figure 3.5 show the resulting p values (and Chi-Squares variation). The missing Journey factors for example are dependent on the Product Description variable $(Chi\ Sq. X^2 = 3.9, p\ value = 0.05)$ as shown by the red line in the Little's MCAR plot in Figure 3.4. This reveals that the journey factors can be derived from a random subset of the Product Description. As such, information from variables related to Product Description are best suited to identify the missing journey factors. A survey for instance identifying the choice of ticket Product Description of passengers (there are about 756 unique products), would be better suited information for use to identify any missing journey factors. In essence, 'Product Description' and variables that are associated with the Product Description are better suited for a journey factor imputation.

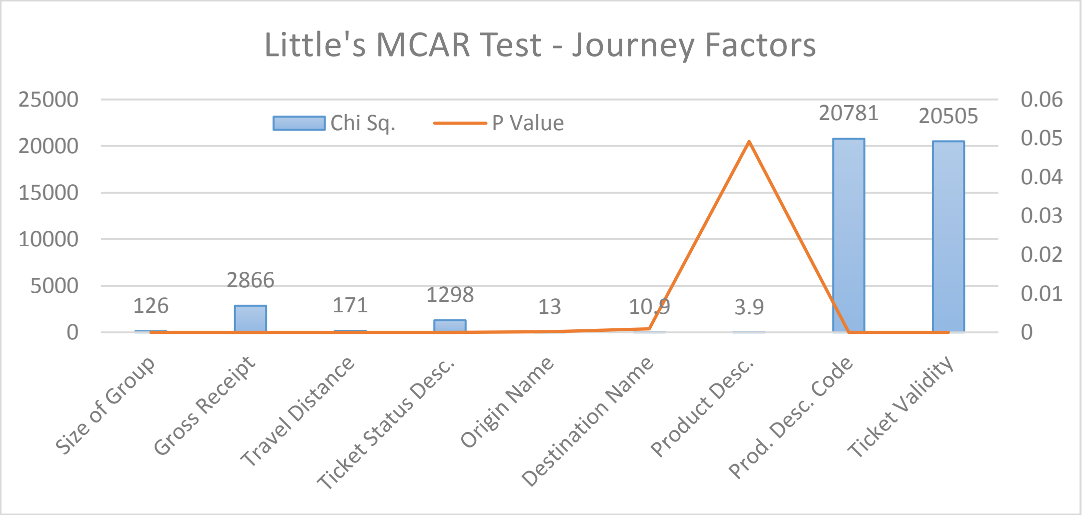


Figure 3.4 MCAR test for missing 'Journey Factors'.

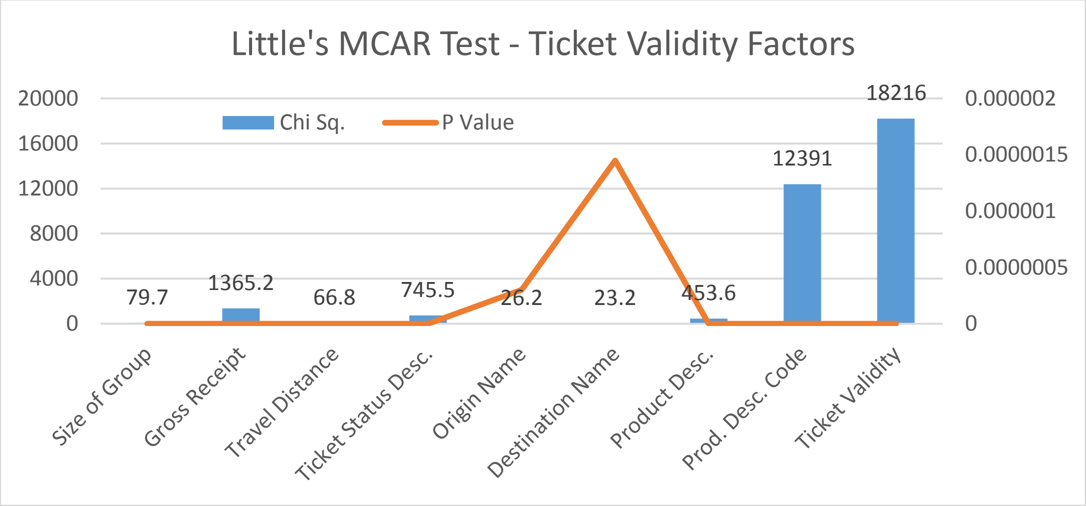


Figure 3.5 MCAR test for missing ‘Ticket Validity’ periods.

The p values $(26 \leq X^2 \leq 18216;\ 0 \leq p \leq 1.45e^{-06})$ in Figure 3.5 for the Ticket Validity periods (offset factors) indicate that the identifiable patterns that exist in missing values within these variables are not significantly related to any variables within the dataset. In such a scenario, additional variables need to be sought and linked-in to explain the difference between the missing and observed values within the ticket validity variable. Such additional variables could be deduced from knowledge of the data generation mechanism. The Grouped Station Indicator variable is binary so t-tests (Aron and Aron, 1994) are more suited to assess its relationship with other quantitative variables (Size of Group, Gross Receipt, Travel Distance and Maximum Validity etc.), excluding the other categorical variables (Origin, Destination, Ticket Description and Codes etc.) which inadvertently may also be important in describing the complex mobility phenomena.

The synopsis of results from Little’s MCAR test is that the Product Description variable better captures random idiosyncrasies that relates to a passengers decision on number of journeys per ticket per day. As such the Product Description as well as additional variables that reveal passenger choice of ticket product description would be better suited to reveal missing journey factors within LENNON. On the other hand, the ticket validity period variable reveals a pattern in its missing values, however Little’s test did not identify any significant other variables within LENNON that could potentially be used to impute missing ticket validity factors. In such a case, an understanding of the data generation mechanism could enable an imputation model to be built (Allison, 2001), or give an indication of appropriate variables that could explain the difference between the observed and missing data values.

3.3 Reconciling datasets

As mentioned in the introduction to this chapter, to facilitate combining disparate datasets, their geographic spatial scales have to be reconciled, as well as the requirement that they should have commensurate variables which enable the datasets to be related. The typical data pre-processing that facilitates the reconciliation of the scale and attributes of the datasets are presented in the following sub-sections. These are the re-zonation process used in this research to reconcile the spatial scale, and for the creation of a new income variable within the Census from alternative available variables. These procedures typically form the first geo-data pre-processing stages, as is the case in this research thesis which develops the framework for harnessing large consumer data.

3.3.1 Boundaries and re-zoning

The geographical boundaries of the UK are complex, multi-layered and non-uniform due to a convoluted evolving history (Byrne, 2000, ONS, 2017). Postcode boundaries were created for the purposes of mail delivery, and have divisions starting from the larger Areas (about 120 of them) covering the UK, to Districts, Sectors (shown on the right of Figure 3.6 for West Yorkshire), and then smaller Postcode Units containing about 15 addresses. On the other hand, the Census boundaries created by the Office for National Statistics (ONS) consists of layers starting from the smallest called the Output Areas (OA's) created from traditional enumeration districts, and made up of about 129 households. The OA's are aggregated to form Lower Layer Super OA's with about 672 households, and then Middle Layer Super OA's (shown on the left of Figure 3.6 for West Yorkshire). Whilst the Census data are spatially scaled to OA boundaries, the NRTS and LENNON ticketing data are scaled to Postcode boundaries[13]. As such, the geography boundary of the Census (LSOA's) was reconciled with that of the NRTS (Postcode Sectors) (Openshaw and Alvanides, 2001). This enables subsequent association to the Postcode Units of the focal LENNON ticketing data.

[13] Within the NRTS, Postcode Unit information exists for each passengers entry and exit train station, along with home address and final destination Postcode Sector. The Census consists of location of usual residence and place of work at 'OA' levels. The LENNON data contains train station of entry and exit at Postcode Unit levels. The re-zonation aims to reconcile the Census OA's with the NRTS Postcode Sector, to enable subsequent association with LENNON Postcode Units.

To explain the sense behind the re-zonation, it is noteworthy to point out that the LENNON ticketing data has information on station entry and exit Postcode Unit, but no information on a representative passenger who might have used the ticket. The NRTS data also has this station entry and exit information at Postcode Unit level. In addition, the NRTS has information on the passenger's home address and final destination at Postcode Sector levels. Thus, the NRTS has representative information on the passengers who might have used the tickets, their entry and exit station, and their home address and final destination. The NRTS also has information on how often a passenger makes a journey in a week, purpose of journey, and income etc., thereby enabling passengers to be broadly classified as commuters and non-commuters. The Census interaction data (also called the flow data) has information at 'OA' and 'LSOA' levels on the home address, final work destination, and mode of travel to work for all passengers within the population. Although there are several modes of travel to work[14], these can be broadly grouped into commute by train and non-commute by train. As seen from Figure 3.1, there are no variables directly linking the LENNON consumer data to the Census information. The association between the LENNON ticketing data and the Census is established via the NRTS. This is the case because the NRTS has the requisite variables to associate with both the LENNON and the Census, as seen from the arrows in Figure 3.1.

The passenger home address and final destination within the NRTS can be linked to that within the Census. Also, the passenger entry and exit station within the NRTS can be linked to the LENNON ticketing data entry and ext stations. The only issue however, is that whilst the NRTS and LENNON are published using the Postcode boundaries (Postcode Unit for station entry and exit, and Postcode Sector for home address and final destination), the Census is published at Output Area (OA or LSOA) levels. As such, the NRTS can be directly linked to the LENNON (as the boundaries are the same). However, a re-zonation would be required to link the NRTS (at Postcode levels) to the Census (at OA level). The re-zonation pixelates the Census OA boundaries, and then re-aggregate the pixels to form Postcode Sector boundaries. This forms the basis of linking the NRTS to the Census flow data, and then associating this to the LENNON ticketing consumer data.

[14] Underground, metro, light rail, tram, train, bus, minibus or coach, taxi, motorcycle, scooter, moped, driving a car or van, passenger in a car or van, bicycle, on foot, other mode of travel, working mainly from home.

The process for re-zoning the Census data from LSOA's boundaries into Postcode Sectors so as to be comparable with the NRTS (Postcode Sectors) was achieved by creating regular square fishnet mesh across the region of interest (West Yorkshire) as shown in Figure 3.6 using purpose developed R-scripts. These variables re-zoned are location of usual residence and place of work by a range of socio-demographic attributes (including age, gender, cars and vans in household, family status, country of birth, hours worked, social grade, economic activity, NS-Sec, industry, and occupation). As the datasets pertain to interaction between origins and destinations, the re-zoning was performed for the origin variables and then repeated for the destination variables. The re-zonation proceeded as follows: In the 2011 Census interaction data, the number of professionals travelling from a usual residence in Middle Super Output Area (MSOA) E02000809 to a place of work in E02000371 is 42. In this category of occupation the number in the reverse direction, i.e. those with usual residence at E02000371 and who work in E02000809 is 29. The re-zonation would consist of dividing the MSOA E02000371 and E02000809 into equal-sized fishnet squares, and assigning a commensurate proportion of the occupation category (i.e. professionals) to each square. The finer the fishnets the more accurate the re-zonation result, as such the fishnets were made 25m squares[15] (finer than that shown in Figure 3.6).

To re-zone the distribution of professionals from MSOA's to Postcode Sectors, the fishnet is applied to the entire study area (West Yorkshire), and the commensurate proportion of professionals in each fishnet square is deduced. The number of professionals in each Postcode Sectors is then estimated by aggregating the fishnet weights that fall within the area of each particular Postcode Sector boundary. This process is repeated for the other categories within the socio-demographic variables. For instance, apart from the 'professionals' category within the occupation socio-demographic attribute, there is a category for 'managers, directors, and senior officials'. This occupation category has 12 people with usual residence in E02000809 who work in E02000371, and 6 people with usual residence in E02000371 who work in E02000809.

[15] The choice of 25m was deemed adequate for this research as even the unusually small Postcode Sectors in the West Yorkshire area are LS155 and LS52 (~8000m^2) with rectangular shapes of about 50m by 150m, which still made a 25m square fishnet adequate.

To minimize the adverse effects of the Modifiable Areal Unit Problem (MAUP)[16], geographical weights were applied to the distribution of proportion of population attributes according to the location of the Postcode Units. This re-zonation spatially scaled the Census attributes to Postcode boundaries, thereby enabling a direct association of Census variables with commensurate variables in the NRTS and then LENNON. In practice the Census interaction data can be structured as many contingency tables, each representing a variable, and the categories within the variable. For instance, the occupation socio-demographic variable would consist of nine contingency tables representing the occupation categories (Managers, directors and senior officials, professional occupations, associate professional and technical occupations, administrative and secretarial occupations, skilled trades' occupations, caring, leisure and other service occupations, sales and customer service occupations, process, plant and machine operatives, and elementary occupations). Each contingency table represents the appropriate flow between the interacting set of MSOA's (Stillwell, 2006).

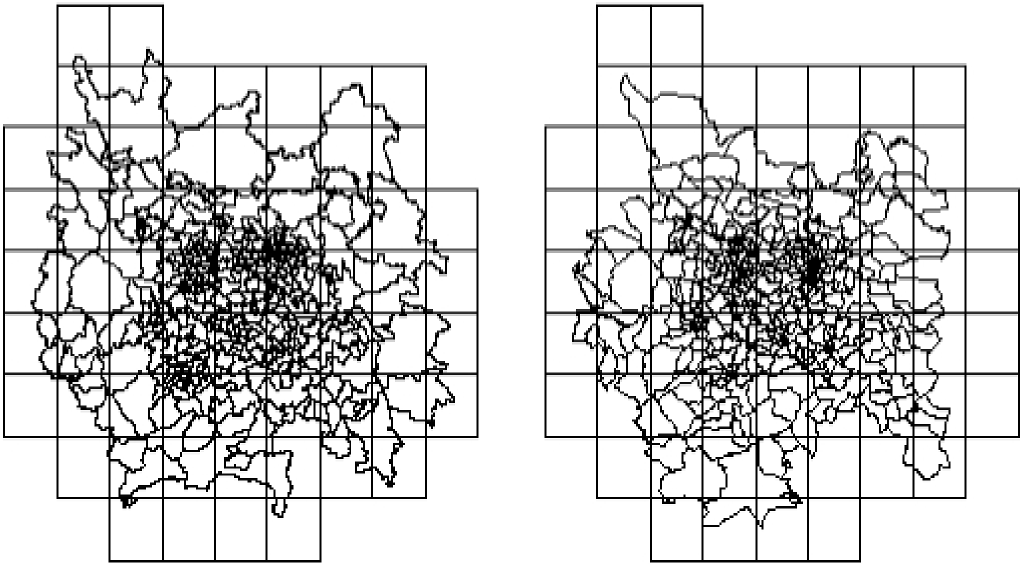


Figure 3.6 MSOA and Postcode Sector re-zonation in West Yorkshire

[16] MAUP represents situations where the result of areal analysis within a geography is dependent on the definition of the geography boundaries. During re-zonation, this phenomena can be alleviated by using point based weights associated with Postcode Units or by offsetting by size of a geographic area

3.3.2 Income for Census

The UK Census does not explicitly contain the 'income' variable to enable direct relations with the 'income' variable in the NRTS. Income is important in travel analysis (Jara-Díaz, 1998, Murakami and Young, 1997), as such a new 'income' variable was created from the 'occupation' variable within the Census. The choice of 'occupation was a result of regressing the Census variables: 'social grade', 'economic activity', 'NS-Sec', and 'industry' and 'occupation', against income from the 2011 ONS Small Area Income Estimates (Henretty, 2011/12). The 'occupation' was chosen as it related (i.e. regressed) best with Small Area Income Estimates, having one of the highest R^2, the lowest Akaike Information Criterion (AIC) (Akaike, 1987), and having the most number of variables with statistically significant p-values. The 'occupation' variable also had an added advantage as it had only nine categories requiring reconciliation with the seven NRTS income categories. The 'NS-Sec[17]' and 'Industry' which had commensurate R^2 values had 15 and 21 variable categories respectively, making them more difficult to reconcile with the NRTS. As shown in Figure 3.7, the personal services and the elementary occupations were combined into one income bracket (income step 2), just as the managers and senior officials were combined with the professional occupations to form the highest income bracket (income step 7). This resulted in a re-categorisation of the 9 categories in the occupation variable, into 7 categories. Figure 3.8 and Figure 3.9 are results showing, in Figure 3.8 a choropleth of the equivalent 'income' variable derived from the 'occupation' variable within the Census. In Figure 3.9 is the choropleth for income derived from the commensurate 'income' variable within the ONS Small Area Income Estimates (SAIE).

On the strength that the 'occupation' variable in the Census and the 'income' variable in the SAIE have higher R^2 values and lower AIC, the 'occupation' is chosen as the closest proxy to the 'income', out of the available socio-demographic attributes. The strength of the association between the 'occupation' and 'income' variables is buttresses and validated by their choropleth plots (Figure 3.8 and Figure 3.9). On the strength of the above, the created 7 occupation categories from the Census are assumed to be the

[17] NS-Sec stands for National Statistics Socio-economic Classification, and is the official socio-economic categorisation adopted in the United Kingdom.

best proxies to the 7 income categories within the NRTS. Just because the Census 'occupation' variable and the NRTS 'income' variable both have 7 categories, this doesn't mean they now correspond. However, the trends of the variables are assumed to be the same as they regress well, although the bands of the categories within both variables may not match exactly. The 'occupation' variable is the best proxy to 'income' from the range of available variables, and this illustrates the nature of considerations that are made in reconciling disparate datasets.

To buttress the point however, basis of the intuition in an expectation of a correspondence between 'occupation' and 'income', is that in the literature (Duncan, 1961, Fuchs, 2004, Soltow, 1960, Wright and Perrone, 1977) the trend of incomes decrease in going from Managers, directors and senior officials, to professional occupations, associate professional and technical occupations, administrative and secretarial occupations, skilled trades occupations, caring, leisure and other service occupations, sales and customer service occupations, process, plant and machine operatives, and elementary occupations. Similarly, the salaries reported in the NRTS 'income' variable are in bands which progressively decrease. As such, although the bands may not be exact, one could serve as a proxy for the other. Despite, this however, it is also noteworthy to point out that care has to be accorded in reporting results from such considerations. In this research, none of the conclusions reported are based in particular on proxy variables.

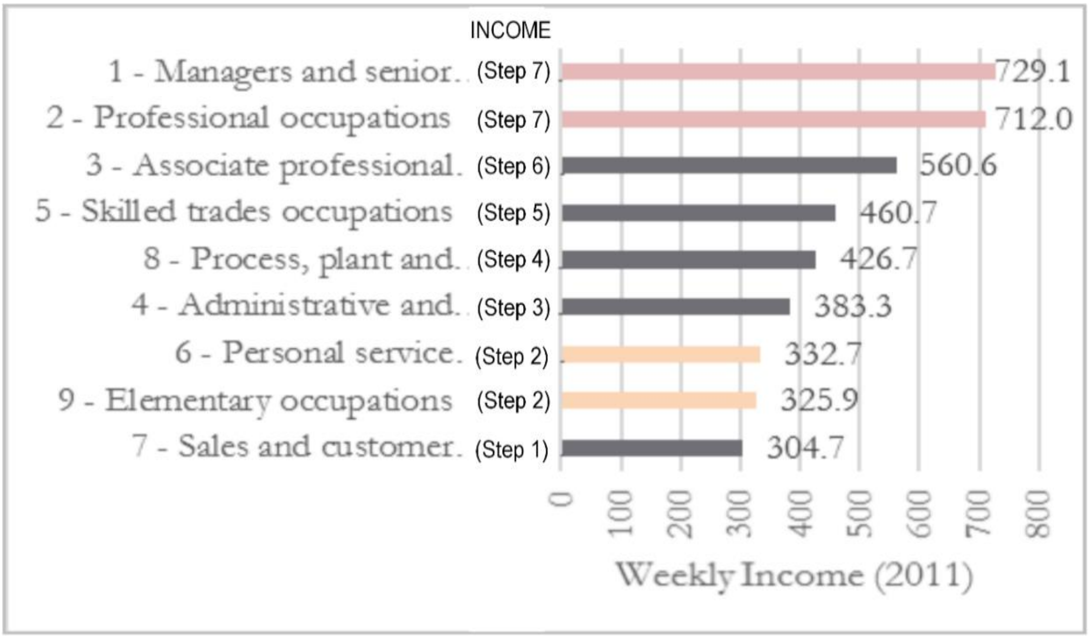


Figure 3.7 Categories of 'Income' derived from 'Occupation' within the 2011 Census.

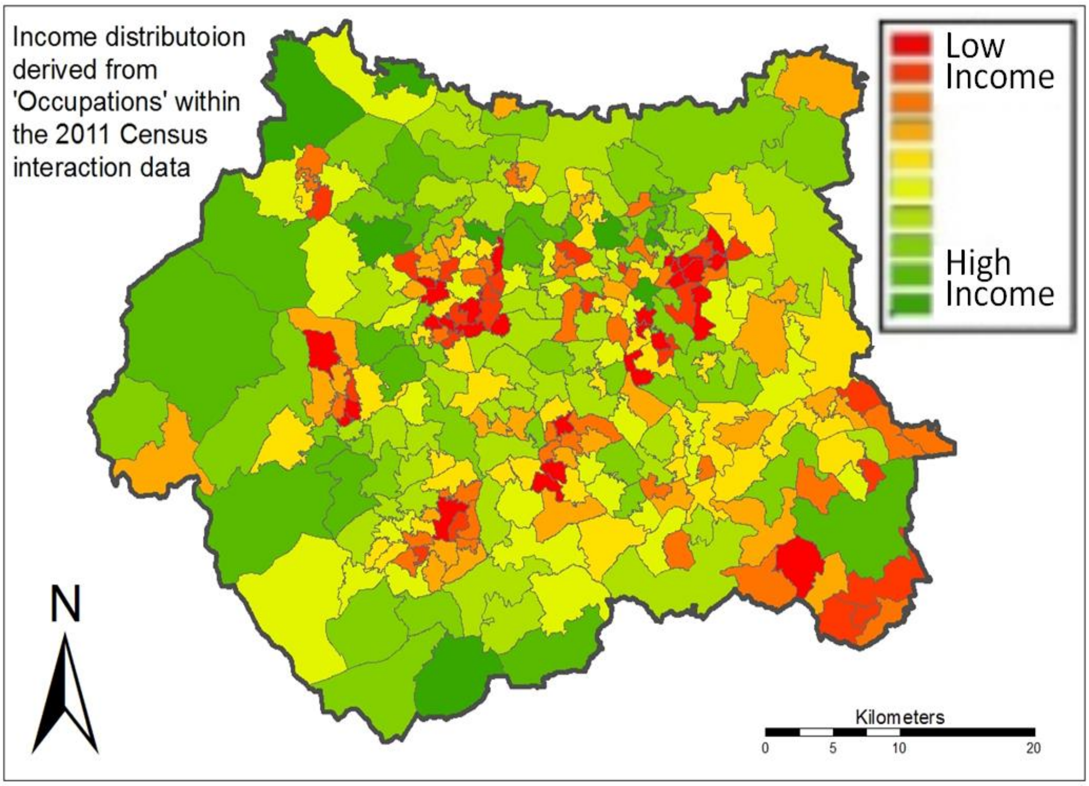

Income distributoion derived from 'Occupations' within the 2011 Census interaction data
Low Income
High Income
N
Kilometers
0
5
10
20

Figure 3.8 Map of ‘Income’ derived from ‘Occupation’ within the 2011 Census.

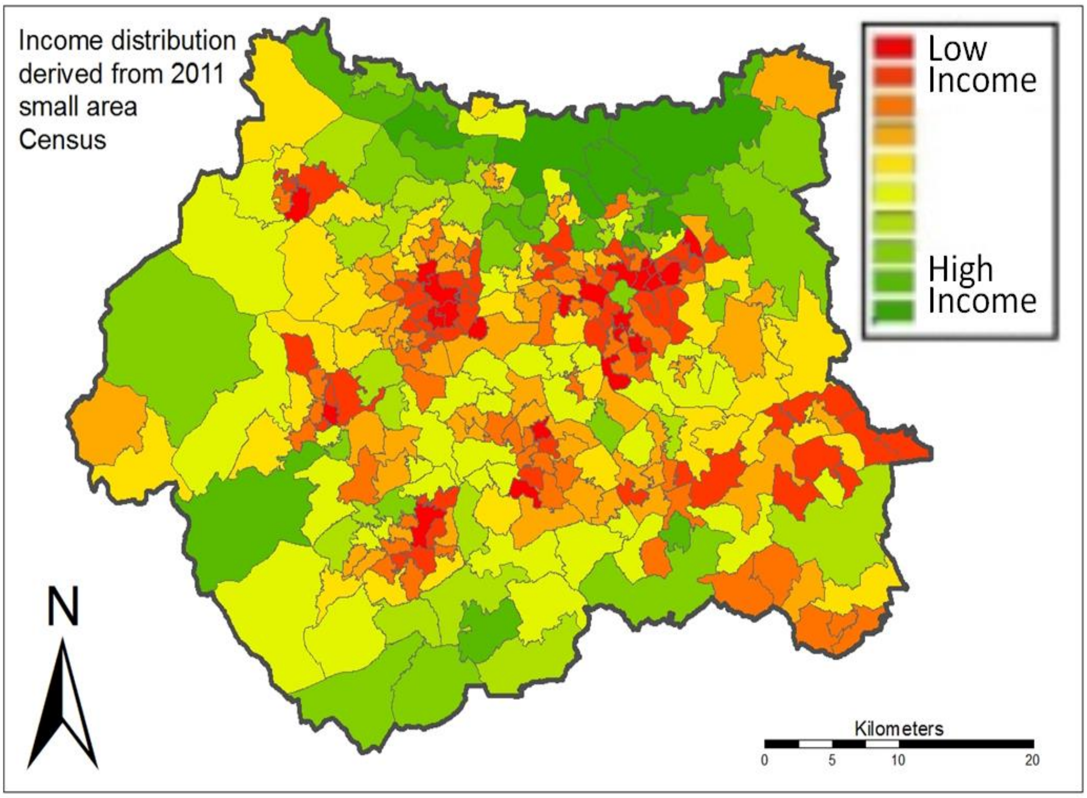


Figure 3.9 Map of ‘Income’ from 2011 Small Area Income Estimates (SAIE).

3.4 Preliminary summary of data

A typical data analytics paradigm would be to summarise important attributes of a dataset, to perform clustering and classification of the variables and values, and then predictive analysis by regression. In practical applications this paradigm is implemented as an iterative development cycle, performing such mobility analysis on the available ‘complete case’ dataset (i.e. the subset of the dataset with all tuples with any missing values removed).’ The LENNON ticketing data without infilling the missing values and linking with other relevant datasets would have the potential to yield results subject to bias, or at best results that are only indicative of the system coefficients. However, such vague indications derived from a complete case analysis could be beneficial in deciding the direction of subsequent investigations when the missing data values are eventually imputed. If subsequent analysis results are markedly qualitatively different, a cogent reason can be sought, adding to the body of knowledge on urban mobility.

In the preliminary analysis of the complete case LENNON data presented below, the first data summary consists of an origin-destination (OD) plot of volumes of passenger flows between the railway stations in the West Yorkshire study area. Then, a visual summary of the relationships between the variables, categories and values within the LENNON dataset are presented using hierarchical classification. The mobility covariates in the rail sector ticketing data are mostly categorical so there are only a limited range of clustering techniques that can usefully be applied. In this preliminary setting, a variant of the K-mode clustering technique is adopted. This is followed by basic predictive regression analysis to give an indication of the parameter values. A complete case analysis of the available data gives an indication (albeit biased) of the relevant parameters of mobility. This will indicate variables that are collinear, such that subsequently when the relevant explanatory predictors are selected, failures due to multicollinearity in redundant covariates can be avoided (Tu et al., 2005). Variable selection methods for datasets with missing values are limited in literature and software implementations are currently not readily available (Garcia et al., 2010). In many practical spatial interaction regression scenarios in the literature (Besag, 1974, Ewing, 1974, Wilson, 1971), the count of passenger groups is set as the dependent response variable, and this value is observed for all ticket sales. The explanatory variables consist of main effects terms (origin and destination) and the interaction distance and cost terms associated with the mobility.

3.4.1 O-D spatial interaction

A 3D origin-destination summary of the LENNON ticketing data is shown in Figure 3.10. The plots show the volume of passengers travelling from an origin train station to a destination station in West Yorkshire. Apart from not having a full repertory of socio-demographic and network attributes associated with railway passengers, the LENNON ticketing data has missing journey factors, ticket validity periods and precise measure of flows to group stations. Notice the fuller volume of passengers in the upper plot reflective of values from current industry infilling of ticket journey factors and assignment of flows to group stations in the LENNON ticketing dataset. The lower-plot in Figure 3.10 is indicative of the lower flow volumes predicted when it is assumed that for instance that all the season tickets have journey factors of unity, and flows to group stations are not assigned. These plot are indicative of the systematic errors that can be introduced when the gaps in ticketing data are not appropriately infilled.

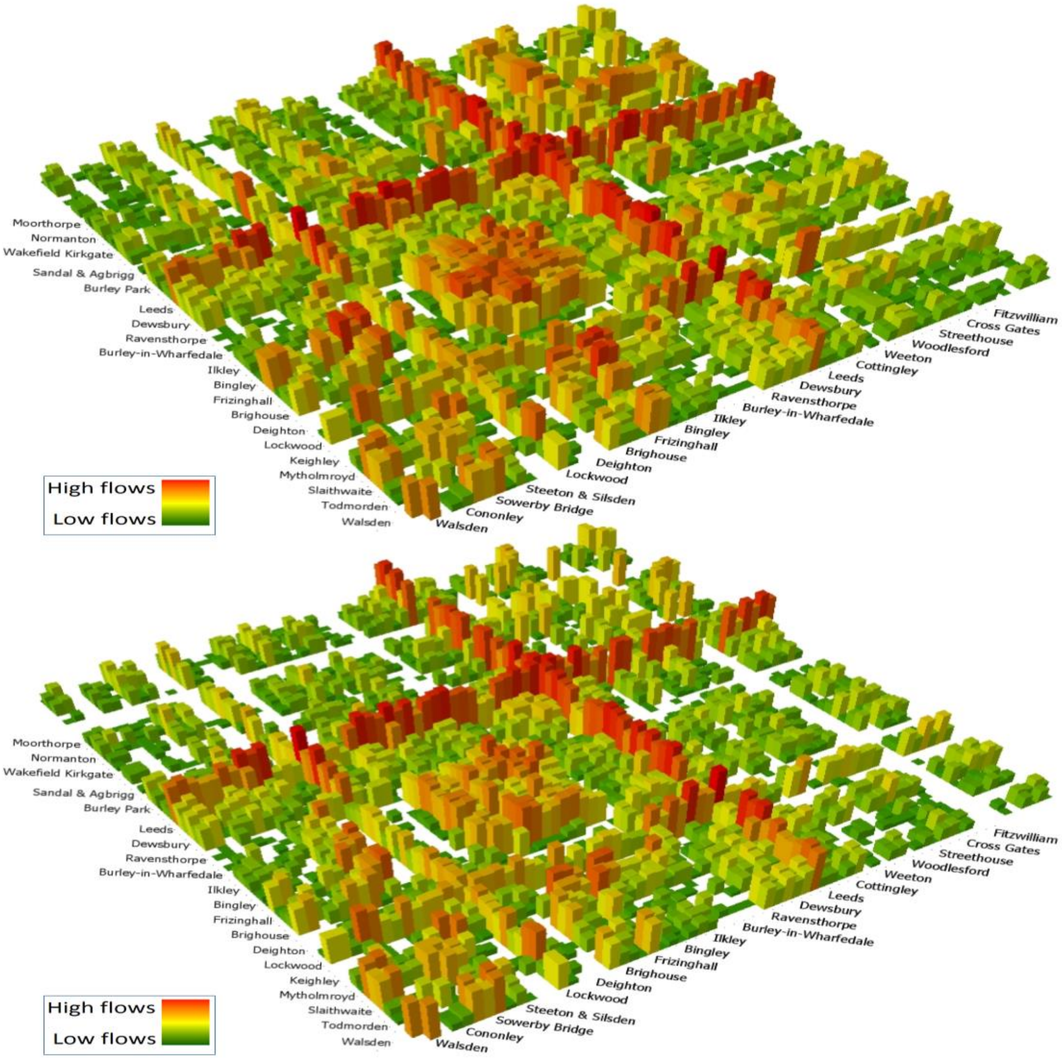


Figure 3.10 3D OD plot (infilled vs missing) for West Yorkshire.

3.4.2 Classification of variables and values

A hierarchical classification was applied to the LENNON tickets to cluster each ticket into a group. As the variables within LENNON are of mixed data types (continuous, ordinal, categorical nominal), the ‘CluMix’ R-software package is applied (Hummel et al., 2017). The particular difficulty in clustering data variables and values containing mixed data types, is the inability to ascribe a distance between variable values once categorical values are involved. Various strategies have been deduced for ascribing a measure of similarity between categorical variables (Galili, 2015, Gower, 1971, Podani, 1999), and some of these are implemented in the ‘CluMix’ R-software package.

The results of the classification are shown in Figure 3.11, and illustrates the relationship between the mixed types of categorical[18] and continuous[19] variables within the dataset. A mixed set of variables that are established in the literature as drivers of mobility interaction have been chosen (i.e. origin, destination, frequency count, travel distance and cost). The dendrogram on the left of the plot shows the relationship between the dependent (‘Freq’) and independent covariates, whilst the dendrogram on the top shows the relationship between the tuples of ticketing values. The heat map enables relationships to be visualised between the hierarchical values in the tree dendrogram and the variables of mobility. This enables a distinction between the relationships between the main effects and the interaction effects.

An inspection of Figure 3.11 reveals a range of preliminary insight into the mobility interaction associated with the railways network in the West Yorkshire study area. Expectedly, lower costs are in general associated with lower travel distances. There are some unusual high distance journeys associated with low costs. These are mainly observed in a select number of flows originating from BIY (Bingley) and destined to LDS (Leeds). Similar observations are seen in a group of flows originating in LDS (Leeds) and

[18] The categorical variables in Figure 3.10 are the origin stations (O-Code) and the destination stations (D-Code). This preliminary summary includes only two categorical variables to start the process of visualising the mobility dynamics. However, a range of additional categorical variables can be included in a systematic study.

[19] The continuous variables investigated in this instance are the travel distance and travel cost. Further continuous variables can be included for a systematic study of the relationship between the covariates.

destined for KEI (Keighley). The dendrogram and heat map hierarchical classification tool can be used to further investigate the data values to, for instance study fare vs distance ratios across all origins and destinations in the study area. Volumes of flows are visually revealed by the area coverage of each colour, and these volumes can be associated with train stations (based on the station colour codes provided). Volumes of particularly long distance journeys are those associated with flows between SON (Steeton & Silsden) to CRG (Cross Gates), and LDS (Leeds) to KEI (Keighley). The hierarchical classification map can also be used to good effect in investigating outliers in the data and the range of values associated with each variable. A limitation with the hierarchical heat map is perhaps the limited variety in the range of colours that can be used as the number of train stations increase such that the distinction between colours becomes blurred. The heat map shown in Figure 3.11 is results for 21 railway stations.

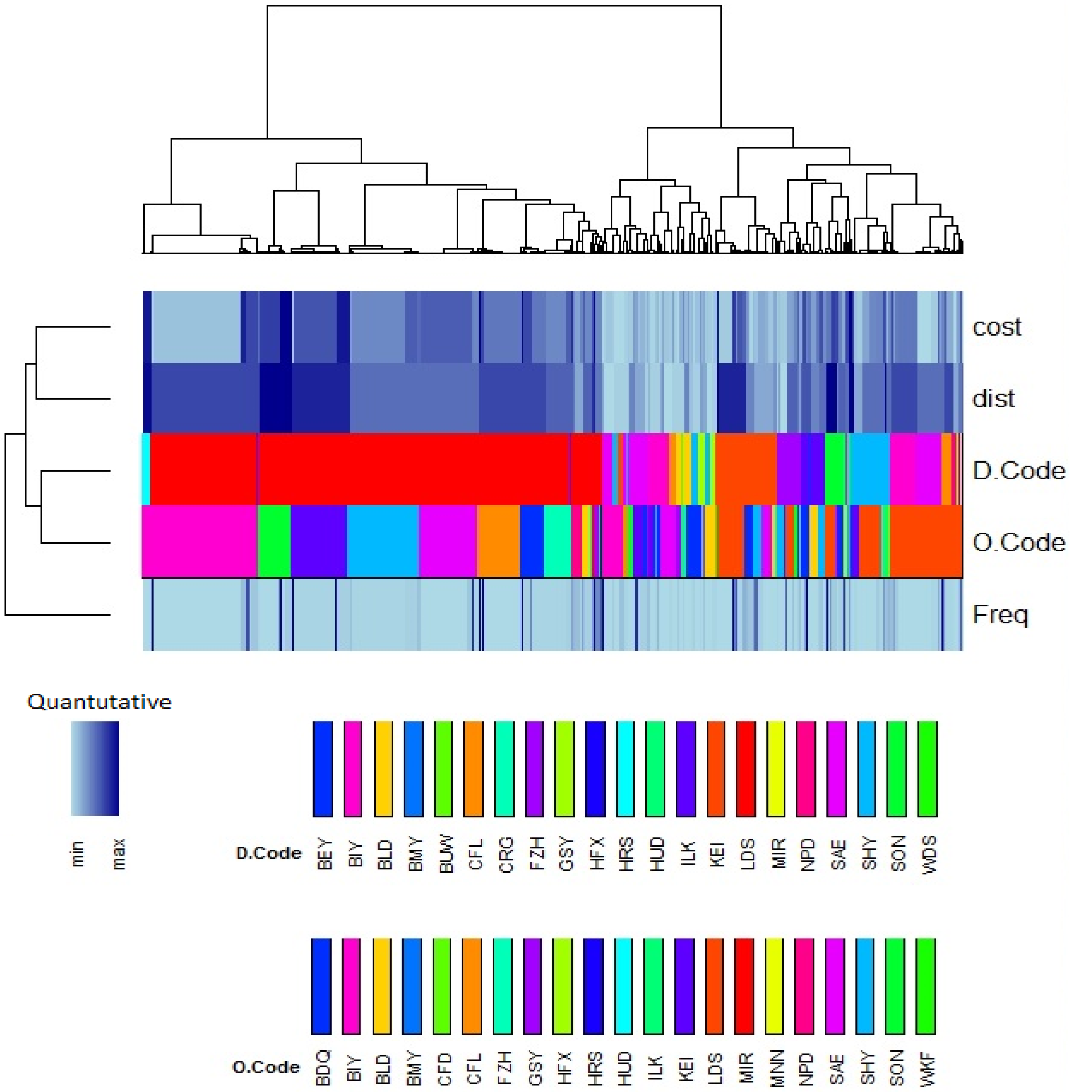


Figure 3.11 Dendrogram of LENNON covariates vs values.

3.4.3 Covariate regression

Once the ticketing data are unravelled exposing the different stages and aspects of the mobility, it is possible to conduct a preliminary exploration of the relationship between the variables of mobility. The gross receipts and travel distance would be influenced by size of group and period of validity of ticket, creating an interaction depicted below

$$Travel\ Distance \propto \frac{Gross\ Receipt}{(Ticket\ Validity)(Size\ of\ Group)}$$

Figure 3.12 illustrates this interaction by relating the gross receipts (log) to the product of ticket validity, size of group and travel distance (log). The offsets in standard (grey), short-term season (blue) and long-term season (red) tickets (with distributions shown in Figure 3.13), are influenced by missing information on the rates of season ticket use within the day and on the lengths of use of flexible period tickets. The systematic vertical shift (offset) in the values are typically resolved by normalising the gross receipt over the journey factors and period of validity of the tickets (for different ticket types, long-, short-term season and standard tickets), had such information been available. The scatter observed in each of the colour strands is indicative of heterogeneity in mobility over the West Yorkshire county in the 4-week period.

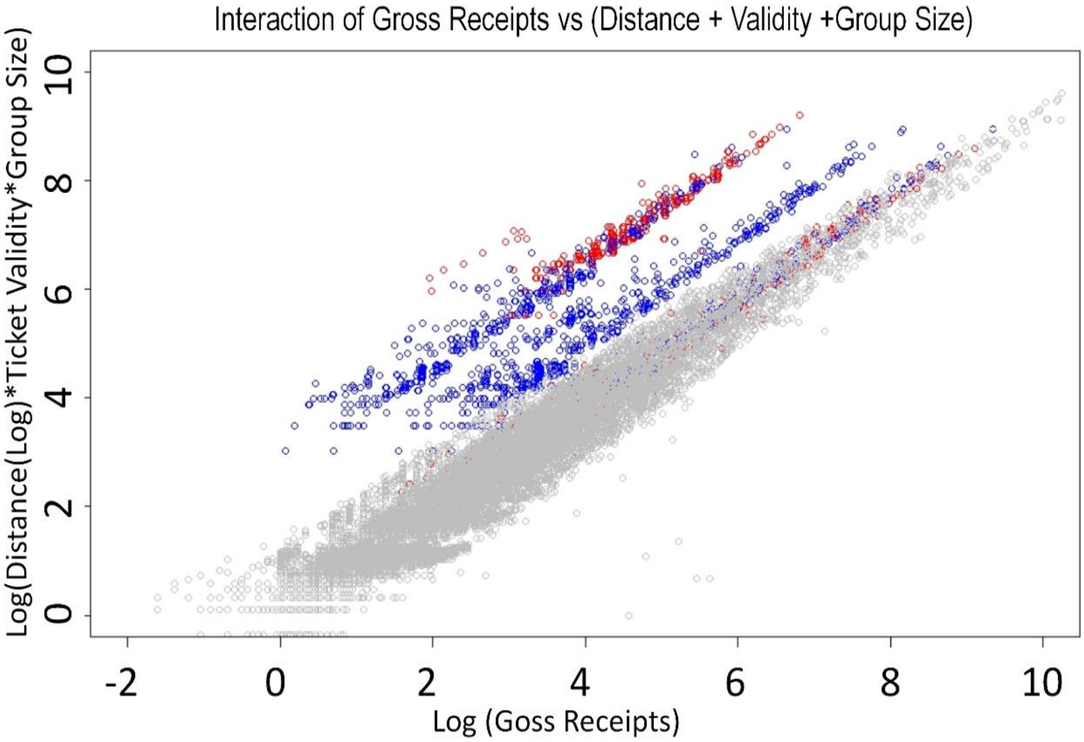


Figure 3.12 Receipts vs Distance, Ticket Validity, and Group Size.

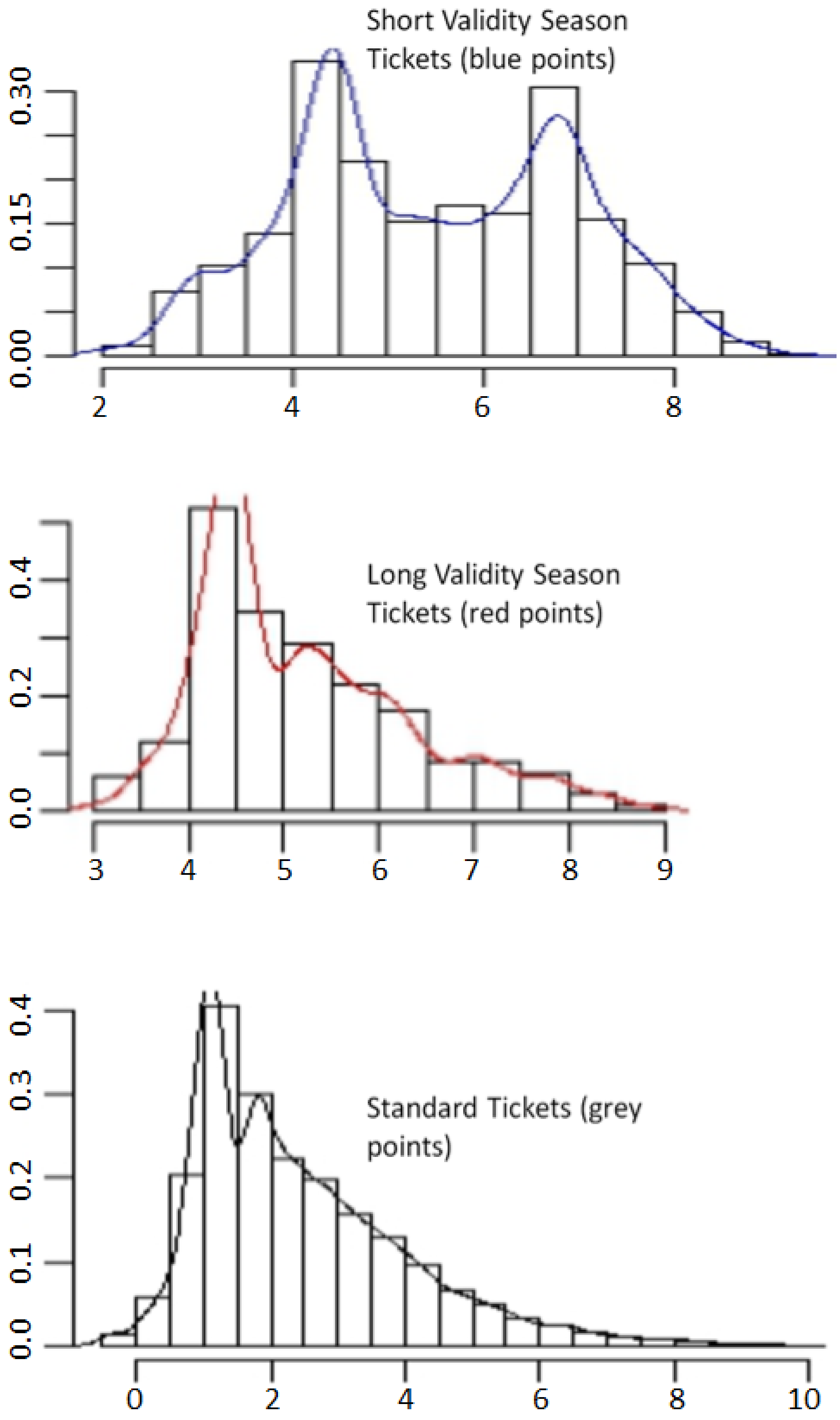


Figure 3.13 Probability density plots for the colour-coded ticket points

A classic predictive regression analysis paradigm for investigating mobility flows of passengers are the spatial interaction models. The histograms in Figure 3.13 are distributions of counts of tickets and are typical of Normal mixtures, Poisson and negative Binomial counts. The homoscedasticity and linear trend of the variables are indicative that a Poisson distribution model for a count outcome variable would be as robust as a negative Binomial model (typically used to account for non-uniform variance across variables in the dataset). Following on from a visual OD matrix shown in Figure 3.10 and the preliminary variable investigation shown in Figure 3.12 and Figure 3.13, a Poisson regression model can be used to model the counts of passengers interacting in mobility between the train stations in West Yorkshire. The explanatory variables consist of main effects passenger origins and destinations, while the interaction effect is captured by the distances between the centroids of the zones. The results of such a model follow on from work widely available in the literature (Batty and Mackie, 1972, Birkin et al., 2010, Dennett, 2012, Fotheringham, 1983, Stillwell, 1978, Wilson, 1971), and yields the coefficients listed in Table 3.2. As is typical of spatial interaction models, the distance parameter is expectedly indicated to be negative, typical of the traditional distance decay.

Table 3.2 Poisson regression results for count of passengers interacting between train station OD pairs.

Covariates (Independent)	**Coefficients (partial)**
Intercept	3.688 *** (0.0868)
Travel Distance	-0.488 *** (0.0015)
Travel Cost	3.152 *** (0.0021)
R-squared	**0.96**
No. of observations	**197,313**

Notes:
Standard errors are reported in parenthesis.
The significance levels at 90%, 95%, 99%, and ~100% level, are respectively indicated as *, **, ***, ****.

3.5 Remarks

It is noteworthy to point out that the preliminary data investigation presented in this chapter are not exhaustive. There are a plethora of other data-point grouping algorithms including but not limited to discriminant analysis, K-means, K-modes, Neural Networks, Random Forests, naïve Bayes, Support Vector Machines etc. Typically data grouping methods that require a training data set are reservedly named classification, otherwise clustering. The difference in the plethora of algorithms also lies in the definition of the group data centre, the distance between groups and in the iteration procedure. The preliminary investigations in this chapter are to illustrate an application suitable for datasets that contain a variety of data-types.

A slightly more elaborate predictive analytics for preliminary investigation of the LENNON ticketing data would be to conduct a hierarchy of logical tests on the covariates using regression trees (Breiman et al., 1984). The trees enable the target variables take on sets of values, indicating only the relevant covariates in the mobility model, thus forming a basis for prediction (Therneau et al., 2017). Regression trees are not explored further in this section as any results will only be indicative, as there are missing values in the covariates. Furthermore, the methodology developed in this thesis for spatial analysis is based on the Bayesian modelling framework.

In discussing the pre-processing stages in this Chapter, the data cleaning up aspects, identification of outliers, visualisation of the ranges of values within each covariate, and correlation between the variables have not been emphasized for various reasons. The less rudimentary cleaning up stages are absorbed within the process of unravelling the consumer ticketing data. Outliers are treated differently in the Bayesian framework adopted for spatial analysis in this thesis. Outliers are readily accommodated as a genuine aspect of the data being modelled, by introducing priors, for instance t-distributions with fat tails to better model outliers, instead of the conventional normal distributions.

This chapter introduced the West Yorkshire study area, and the associated consumer data. The gaps in the consumer data are discussed, and the formalisation of the concepts that enable the plugging of the data gaps are presented. The following chapters in Part 2 of the thesis present the set of three complementary methodologies to harness the consumer data. Once the data are harnessed, it becomes useful for objective mobility analysis.

PART 2

This part presents the core aspects of the thesis. In this part a set of powerful concerted complementary methodologies are developed, which together for the framework that enables big consumer data to be harnessed for use in urban mobility and movement patterns analysis.

Chapter 4
SPATIAL MICROSIMULATION

This chapter presents spatial microsimulation, the first of three concerted methodologies used to harness consumer datasets. Consumer data are typically sensed for a particular purpose, and do not have the broad spectrum of attributes for wider application. Consumer data are also not like traditional measured stated preference surveys from a representative wider population, instead, they are revealed preferences from consumers of a particular service thereby forming a skewed subset of the wider population. Further, apart from unobserved attributes, consumer data tend to have missing values as the sensors do not always capture the full context of the data generation process. Finally, certain idiosyncratic behaviour of passengers (like daily rates of season ticket use, and the particular choice of exit station when the ticket is valid for travel to a group of stations) are typically unobserved. These issues with novel consumer datasets form the gaps in the data, and these gaps have to be plugged to create a harnessed and requisite dataset for mobility analysis. The aim of this chapter is to address the first set of gaps in consumer data, i.e. by increasing the spectrum of data attributes for a case when the consumer data (like the ticketing data) are a skewed subset of the wider population (like the Census). The objective is to validate current spatial microsimulation methods for application to typical consumer datasets. To achieve this, experiments are conducted to ascertain the behaviour of a range of spatial microsimulation methods when the input data suffers the issues associated with typical consumer data.

4.1 Literature on spatial microsimulation

In the absence of an attribute-rich, comprehensive and representative population, a simulated population is necessary for the analysis of mobility on the railways. Consumer datasets tend to have comprehensive specific coverage, but are not representative of the entire population. The heterogeneity in these datasets can be better harnessed by integrating them with other relevant measured stated preference surveys which are designed to be random samples representative of the population. In this chapter both deterministic and stochastic spatial microsimulation strategies are discussed for combining various datasets, creating a representative micro-mobility population of railway passengers embedded in the wider population within a geographic region (West Yorkshire, UK). As discussed in Chapter 2, a deterministic methodology particularly suited to adjust skewed rail-mobility consumer data is developed, combining information on all rail tickets sold in the UK, with the 2011 Census commute to work data, and a National Rail Travel Survey (NRTS), yielding a representative micro-level population. The deterministic strategy using multi-dimensional iterative proportional fitting (m-IPF) is presented in a practice-oriented way to enable reproduction of the methodology, highlight nuances, precautions, and associated advantages of the methodology. The simulated population created includes weights which represent the probability density of the railway passenger population.

As set forth in chapters 1 and 2, the skewed nature of consumer data poses a challenge in integrating with other relevant datasets from measured stated preference surveys which are designed to be random samples representative of the population. Traditional spatial microsimulation strategies tend to assume that the disparity is normal between the datasets being combined (Cox and Snell, 1968, Robert, 2004). As such, this chapter validates the range of established spatial microsimulation strategies for use to adjust skewed 'big data[20]' for consistency with survey data and established theory.

Early Census exigencies (see Section 2.4) meant that despite complete counts of the populace, only a small sample of anonymised records (SAR) of cross-tabulations of all individual attributes were released, alongside comprehensive sets of aggregate tables typically limited to only three variables per table. Despite the anonymised data and aggregation in respect of individual privacy, population geographers realise that comprehensive and disaggregated records would enable the construction of more detailed pictures of geographic phenomena in nature. The need to create such micro-data led geographers to develop the field of study of spatial microsimulation (Birkin and Clarke, 1988). The first application of spatial microsimulation (population synthesis) is widely attributed to Deming and Stephan (1940), who applied the method to the USA Census. A synthetic population was formed by estimating an optimal combination of multiples of individuals in a sample of anonymised records that will best fit the aggregate values defined in Census tables, using the Lagrange multiplier method (Bertsekas, 2014).

[20] Big data is a term coined to describe the range of novel large consumer datasets that are increasingly readily available from sensors of consumers of digital services. Refer to Page 4 for another definition.

The Lagrange multiplier was developed in the 18th Century (Clarke, 1990, Gay, 1966, Suzuki, 2005) to solve optimisation problems encompassed by a constraint. The conditions imposed on the objective function requiring optimisation can take on a number of forms including least squares, maximum likelihood and chi-squares etc. (Deming and Stephan, 1940). As discussed in chapter 2, with the development of computers and numerical analysis, instead of solving the linear equations derived from the objective function, iterative methods gradually converging towards an optimum were proposed (Fletcher, 2013, Kelley, 1999), and it became increasingly more efficient to resort to these iterative proportional fitting (IPF) strategies.

In population geography and demography, IPF came to prominence relatively recently with the range of policy relevant applications and solutions proffered. Birkin and Clarke (1989) used IPF to simulate individual and household incomes at small area levels, Rees (1994) used IPF to project age and gender structure of urban areas, and Ballas et al. (2007) addressed in detail the use of spatial microsimulation as a framework for decision support for policy analysis, and Ballas and Clarke (2001) assessed the impact of aggregate national policies within segregate local communes. The IPF methods have improved with burgeoning use, and while earlier effort concentrated on developing the microsimulation steps on different platforms, the advent of more able computers and suites of statistical software packages like R and R-studio (RStudioTeam, 2016) have enabled an automation of the processes. Effort shifted to concerns of numerical stability and propagation of errors in IPF methods (Birkin and Clarke, 1995, Wong, 1992). Further concerns relate to converting fractional values for simulated individuals to integer counts of whole individuals, and in developing strategies for internally and externally validating IPF results (Lovelace and Ballas, 2013, Upton, 1985). Spatial microsimulation is now considered a mature application (Lomax and Norman, 2016), re-visiting questions concerning the reliability and confidence in unconstrained attributes that are reproduced by the IPF process (Birkin and Clarke, 2011). As such, how accurately are the distributions of those variables not included in the spatial microsimulation replicated? Further unanswered questions concern whether a large number of benchmark constraining variables generally translate to better microsimulation results (Smith et al., 2009), the effect of the quality of the seed sample, and the sensitivity of the different spatial microsimulation strategies to sample ratios (Tanton, 2014). These are some of the questions answered in this chapter in addressing the practicality of spatial microsimulation methods.

Apart from deterministic (IPF) spatial microsimulation strategies which yield the same result on repeat, the alternative strategies are stochastic, typically Markov chain Monte Carlo (MCMC) based methods (Metropolis et al., 1953). The MCMC algorithms are a most efficient method for converging to solutions of intractable complex problems, likened to constrained optimisation of the class of spatial microsimulation (Asmussen and Glynn, 2007, Brooks et al., 2011). Population geographers have developed a range of MCMC variants to complement traditional deterministic IPF.

Some of these methods (already discussed in Section 2.4.2) have been branded simulated annealing and hill-climbing, and have been compared with the IPF method (Harland et al., 2012, Kurban et al., 2011). Another MCMC based strategy the genetic algorithm was used to reconcile observed and estimated network traffic flows (Dimitriou et al., 2006). A further stochastic strategy for spatial microsimulation is the synthetic reconstruction (Birkin and Clarke, 1988, Birkin et al., 2006), which used the Sample of Anonymized Records (SARs) from the Office for National Statistics (ONS). More recently, further sets of MCMC based population synthesis methods have been developed in the transportation sector. One is the simulation based strategy which adopts Gibbs sampling (Farooq et al., 2013), and another adopts generative methods and is based on the rejection sampling (Sun and Erath, 2015).

As mentioned earlier (in Chapter 1 and Chapter 2), novel consumer datasets and so-called big data, tend to be a skewed subset of the population. As a result of this skew, the assumption of normally distributed errors between the target[21] population (also referred to a posterior) and the seed data (also referred to as a proposal[22] distribution) no longer holds (Deming and Stephan, 1940, Rubin, 1976). As big data tends to suffer sample bias (Kitchin, 2014), methodologies developed which assume normal errors between datasets need to be reviewed prior to application to these novel consumer datasets. As a result, spatial microsimulation strategies traditionally set up for representative population seed samples, need to be validated for application to skewed consumer seed data. For the

[21] Target distribution refers to the distribution of population being sought. Posterior distribution is the result of a stochastic simulation. The aim is that the posterior distributions converges to the target distribution

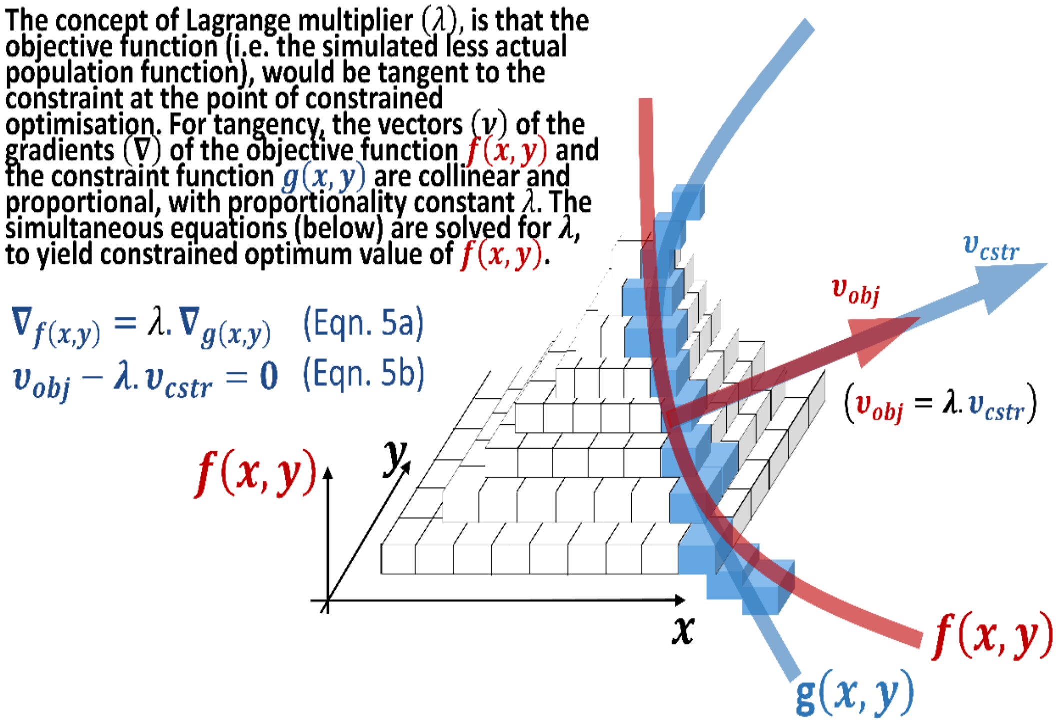
[22] Proposal distribution is the seed sample which is replicated to simulate a representative population.

deterministic IPF then, the promise in application to skewed data lies in establishing the behaviour of the iterations that define the procedure. For the stochastic procedures typically based on variants of the MCMC (Richey, 2010, Rosenthal, 2011), the promise in application to skewed datasets lies in the ability of the sampling and optimization chains to converge to the target population, given the skewed seed. In this chapter, the behaviour of deterministic and stochastic spatial microsimulation procedures are investigated, to understand the nature of the methodologies in anticipation of application to the railway consumer data from the West Yorkshire study area.

4.2 The Lagrange multiplier

Spatial microsimulation methods are described as ways of solving particular optimisation problems, and the Lagrange multiplier is one such deterministic method (Sun and Yuan, 2006). In spatial microsimulation the typical aim is to create a best possible (i.e. optimal) synthetic representative population of a geographic zone, limited (i.e. constrained) such that the aggregate volume and characteristics of the zone are fulfilled. The classic solution strategy for constrained optimisation problems have been to use calculus (Bertsekas, 2014, Suzuki, 2005). In scenarios of multi variables, partial differential equations enable us assess the effect of a change in one variable when all the others are kept constant, enabling the total effect of partial changes in each variable. The optimal solution is thus inferred when the total effect of changes in the variables is tangent to the constraint of the problem (Bertsekas, 2014).

The premise of the Lagrange multiplier is that the objective function (the function being optimized) would be tangent to the constraint variable at optimum. The vector of this tangent would be collinear with the constraint, and equivalent subject to a proportionality constant λ (where λ is defined as the Lagrange multiplier). This is illustrated in Figure 4.1. At the constrained optimum, the objective function and the constraint function would be tangent and the vector gradients of the functions (represented by the red and blue arrows respectively) would be collinear and proportional, with proportionality constant λ the Lagrange multiplier. The result in Equation 5a, 5b (embedded in Figure 4.1) are solved for λ, yielding the other optimisation parameters (Bertsekas, 2014, Deming and Stephan, 1940, Everett III, 1963). The Lagrange multiplier method is discussed here for historical reasons but, as Section 4.5.1 will discuss, are intractable in situations with thousands of individuals (as is the case in this study).



The concept of Lagrange multiplier (λ), is that the objective function (i.e. the simulated less actual population function), would be tangent to the constraint at the point of constrained optimisation. For tangency, the vectors (v) of the gradients (∇) of the objective function $f(x,y)$ and the constraint function $g(x,y)$ are collinear and proportional, with proportionality constant λ. The simultaneous equations (below) are solved for λ, to yield constrained optimum value of $f(x,y)$.

$$\nabla_{f(x,y)} = \lambda . \nabla_{g(x,y)} \quad \text{(Eqn. 5a)}$$

$$v_{obj} - \lambda . v_{cstr} = 0 \quad \text{(Eqn. 5b)}$$

Figure 4.1 Constrained optimisation using Lagrange Multiplier.

4.3 Deterministic and stochastic strategies

The methods used for spatial microsimulation can be broadly classed as deterministic or stochastic strategies. The particular name given to the deterministic method in turn depends on the description of the objective function – least squares and chi-squares difference, maximum normal likelihood all refer to the conditions imposed on the objective function. Other names for methods which used the Lagrange multiplier method are the Entropy maximisation which is a solution concept derived from the thermodynamics discipline, but which has wide context and out-of-discipline application (Batty and Mackie, 1972, Wilson, 2010). The Entropy maximisation modelling concept is premised on the maximum information being derived by models with least amount of pre-conceptions about the solution. A general constraint imposed on the outcome of these models yields an exponential power law functional form which has wide application in modelling many phenomena in science (Shore and Johnson, 1980, Wilson, 1969). In preference to using the Lagrange method and solving the normal expressions for the objective functions, the iterative proportional fitting (IPF) algorithm is typically used in practice and has been shown to be an algebraic simple, computationally fast routine which is also robust to input errors (Bishop et al., 2007, Deming and Stephan, 1940, Fienberg, 1970).

In this chapter, the IPF[23] strategies are explored for application to deterministic spatial microsimulation, in a bid to identify the behaviour of the different solution strategies for subsequent application to simulating a micro-population of railway passengers. The alternative to the deterministic strategies are the stochastic ones, which have been called combinatorial optimisation and reweighting strategies. Strictly however, the deterministic (IPF) strategies also produce weights for the individual-seed (Lovelace and Ballas, 2013), the contrast is perhaps in the integer weights produced by stochastic strategies and the fractional weights for deterministic strategies. In optimisation theory (Bertsimas, 1988, Mühlenbein et al., 1988), all problems with discrete variables are categorised as combinatorial optimization, so strictly both our deterministic and stochastic strategies can be classed as combinatorial optimization as well as re-weighting strategies. Here, deterministic and stochastic microsimulations strategies are the preferred nomenclature, and not re-weighting or combinatorial optimisation.

As distinct from deterministic strategies which iteratively improve the entire seed sample, the stochastic strategies are typically variants of the MCMC strategy, which involve two stages - an exploratory and an optimisation stage. The exploratory stage involves randomly or systematically searching the range of values within the seed sample to find potential replacement samples that may improve a proposal density, and an optimisation stage to decide on implementation of any improvements. Two sampling strategies are predominant in the MCMC methods, the Metropolis-Hastings and the Gibbs sampler, used to conceive ways of effectively sampling from the target distribution (also called the posterior). In the Metropolis-Hastings sampling (from which the Simulated-Annealing and Hill-Climbing strategies are derived (Kavroudakis, 2015, Williamson et al., 1993)), proposal seed samples are taken, and these samples are recursively improved by repeated random samples each time comparing with the target and deciding a criteria for acceptance. The Gibbs sampler is distinct as it is premised on being able to sample the full joint posterior distribution (Chib, 1995, Farooq and Bierlaire, 2017). In practical application where Gibbs sampler is used for spatial microsimulation, a parametric model of the system being simulated is created combining variables from the seed sample with partial views of the

[23] The notable shortcoming of IPF is that it does not converge if there are non-common marginal values or non-representative seed elements for the benchmark constraints (Saito, 1992, Barthelemy and Suesse, 2016). These are issues explored in this chapter.

full joint posterior within the target cross tables. The parameters of this model are iteratively and sequentially improved conditional on the last set of improvements until convergence. The converged parameter values are used as a basis for estimating posterior predictive distributions which form the simulated population (Farooq and Bierlaire, 2017).

In this section, the literature review on the range of deterministic and stochastic spatial microsimulation methods has been presented. This review serves as a precursor to the spatial microsimulation methodology presented in the next section. The concepts and procedures of the range of spatial microsimulation methodologies are presented next.

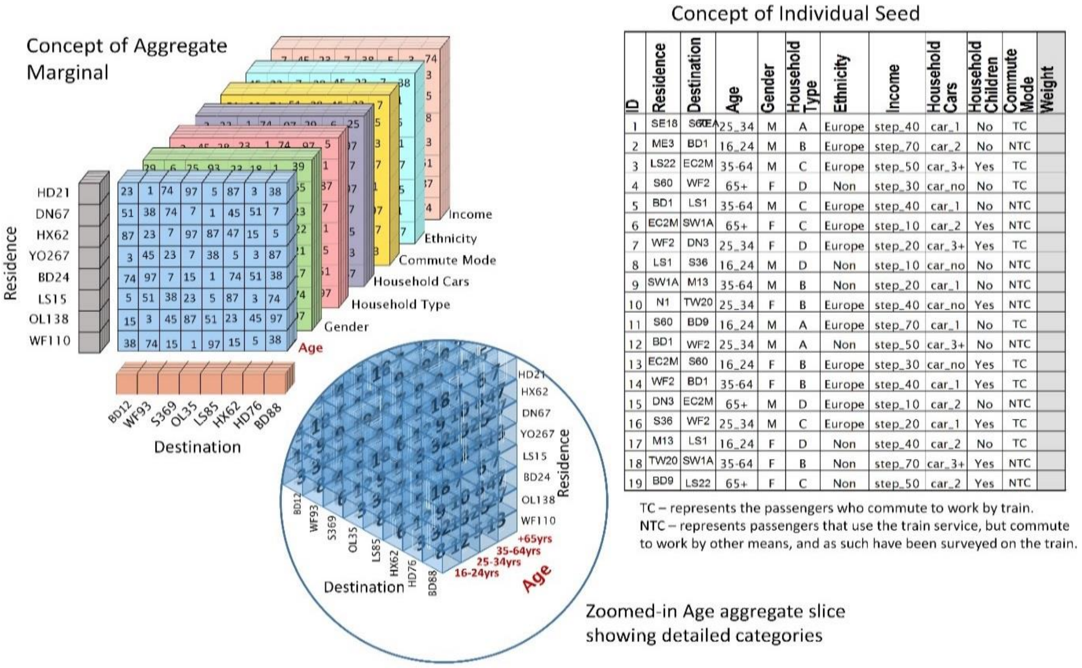
4.4 Methodology for spatial-microsimulation

The methodology for deterministic and stochastic spatial microsimulation is presented below. Concepts applicable to both strategies are first presented, and then particular procedures adopted for the deterministic and stochastic methods are presented separately. The IPF strategy is the more popular of the deterministic spatial microsimulation strategies (Tanton, 2014), which has evolved over the years into a multi-dimensional application better suited for combining multi-facetted datasets (Barthelemy et al., 2016). Other deterministic strategies based on the Lagrange multiplier method include the least squares, maximum likelihood and chi-squares methods (Deming and Stephan, 1940). A detailed exposition on deterministic spatial microsimulation methods can be found in the literature (Tanton and Edwards, 2012). There are four generally known MCMC based stochastic strategies namely the hill climbing, simulated annealing, genetic algorithm, and simulation-based population synthesis. The hill climbing (HC) method which is akin to the MCMC rejection sampling (Gilks et al., 1995), the simulated annealing (SA) which is akin to the MCMC Metropolis-Hastings algorithm (Chib and Greenberg, 1995), the genetic algorithm (GA) which mimics the optimal evolution process of nature (Davis and Principe, 1993), and the simulation-based population synthesis (S-B) which is based on the Gibbs sampler (Farooq and Bierlaire, 2017). These deterministic and stochastic methods are briefly discussed and subsequently assessed for application to skewed datasets typical of the NRTS and LENNON ticketing data from the rail sector. A more detailed treatment of HC and SA algorithms tailored to population geography are available in literature (Ballas et al., 2005, Williamson et al., 1998).

4.4.1 Concepts behind methods

The concept of spatial microsimulation is embodied in the illustration presented in Figure 4.2. In practical applications in population and transport geography, the right table depicts a representative population sample, or a subset of the population sample, listing a number of individuals on the rows (tuples) and attributes of these individuals within the column-fields. In the literature, this table is referred to as the seed, the sample, individual-level data, and survey or simply as the seed. The left rectangular slab represents the structural form of zonal population attributes, usually aggregated and cross-tabulated. The dimensions of the aggregate structure could represent a zonal attribute and the categories therein. The vertical grey pillar (on its own) for example represents the 'Residence' variable and the geographic zone categories therein. If values were included in each cube that made up the vertical grey pillar, these would be the populations associated with each residential zone. A similar description follows for the horizontal brown pillar which (on its own) represents the 'Destination' attributes and the zone categories that make up the destinations. The vertical pillar represents a one-dimensional $(1D)$ aggregate constraint, just as the horizontal pillar also forms an $1D$ aggregate constraint. If the vertical or horizontal pillars are further sliced length-wise, it would become a $2D$ constraint; with the second dimension representing say the age variable, sliced into the categories (16-24yrs, 25-34yrs etc.). The slabs in Figure 4.2 form $2D$ constraints, and in this instance the blue slab in front represents an aggregate array cross-tabulation of 'Residence' versus 'Destination'. The blue slab is further sliced along the vertical axis, creating the third 'Age' variable, with categories therein. In such an instance, the aggregate constraint would be an $3D$ array. The categories making up these 'Residence', 'Destination' and 'Age' variables are illustrated in the blown up section (highlighted in the oval section).

As seen in the blown up section, for passengers residing in Postcode WF110 and working at destinations in Postcode BD88, there are populations of 8, 12, 5 and 13 of ages 16-24yrs, 25-34yrs, 35-64yrs and over 65yrs respectively. The subsequent slabs coloured green, pink, purple, etc. also represent $3D$ aggregate constraints. As seen, the variables involved in each slab are 'Residence', 'Destination', and another demographic attribute. Each of the variables are further sub-divided into categories (typical of the blown up 'Age' variable in the circle). The 2011 Census interaction data consists of several such 3D arrays (like those in Figure 4.2) made up of aggregates for location of usual residence and place of work for a range of socio-demographic attributes (age, gender, income, mode of commute etc.). Zonal population aggregates of cross-table data are usually simply referred to as aggregate, marginal, constraint, count, target, posterior distribution or Census (as its structural form is typical of published Census counts). Typically in practice, data available to population geographers would consist of aggregate and seed datasets, and spatial microsimulation enables the combination of these to yield a representative individual-level dataset. If the microsimulation process is successful, then the representative micro-level population created would inherit the attributes of the aggregate and seed.



	BD12	WF93	S369	OL35	LS85	HX62	HD76	BD88
HD21	23	1	74	97	5	87	3	38
DN67	51	38	74	7	1	45	51	7
HX62	87	23	7	97	87	47	15	5
YO267	3	45	23	7	38	5	3	87
BD24	74	97	7	15	1	74	51	38
LS15	5	51	38	23	5	87	3	74
OL138	15	3	45	87	51	23	45	97
WF110	38	74	15	1	97	15	5	38

Concept of Individual Seed

ID	Residence	Destination	Age	Gender	Household Type	Ethnicity	Income	Household Cars	Household Children	Commute Mode	Weight
1	SE18	S6TEA	25_34	M	A	Europe	step_40	car_1	No	TC	
2	ME3	BD1	16_24	M	B	Europe	step_70	car_2	No	NTC	
3	LS22	EC2M	35-64	M	C	Europe	step_50	car_3+	Yes	TC	
4	S60	WF2	65+	F	D	Non	step_30	car_no	No	TC	
5	BD1	LS1	35-64	M	C	Europe	step_40	car_1	No	NTC	
6	EC2M	SW1A	65+	F	C	Europe	step_10	car_2	Yes	NTC	
7	WF2	DN3	25_34	F	D	Europe	step_20	car_3+	Yes	TC	
8	LS1	S36	16_24	M	D	Non	step_10	car_no	No	NTC	
9	SW1A	M13	35-64	M	B	Non	step_20	car_1	No	NTC	
10	N1	TW20	25_34	F	B	Europe	step_40	car_no	Yes	NTC	
11	S60	BD9	16_24	M	A	Europe	step_70	car_1	No	TC	
12	BD1	WF2	25_34	M	A	Non	step_50	car_3+	No	NTC	
13	EC2M	S60	16_24	F	B	Europe	step_30	car_no	Yes	TC	
14	WF2	BD1	35-64	F	B	Europe	step_40	car_1	Yes	TC	
15	DN3	EC2M	65+	M	D	Europe	step_10	car_2	No	NTC	
16	S36	WF2	25_34	M	C	Europe	step_20	car_1	Yes	TC	
17	M13	LS1	16_24	F	D	Non	step_40	car_2	No	TC	
18	TW20	SW1A	35-64	F	B	Non	step_70	car_3+	Yes	NTC	
19	BD9	LS22	65+	F	C	Non	step_50	car_2	Yes	NTC	

TC – represents the passengers who commute to work by train.
NTC – represents passengers that use the train service, but commute to work by other means, and as such have been surveyed on the train.

Figure 4.2 Data structures used in spatial microsimulation.

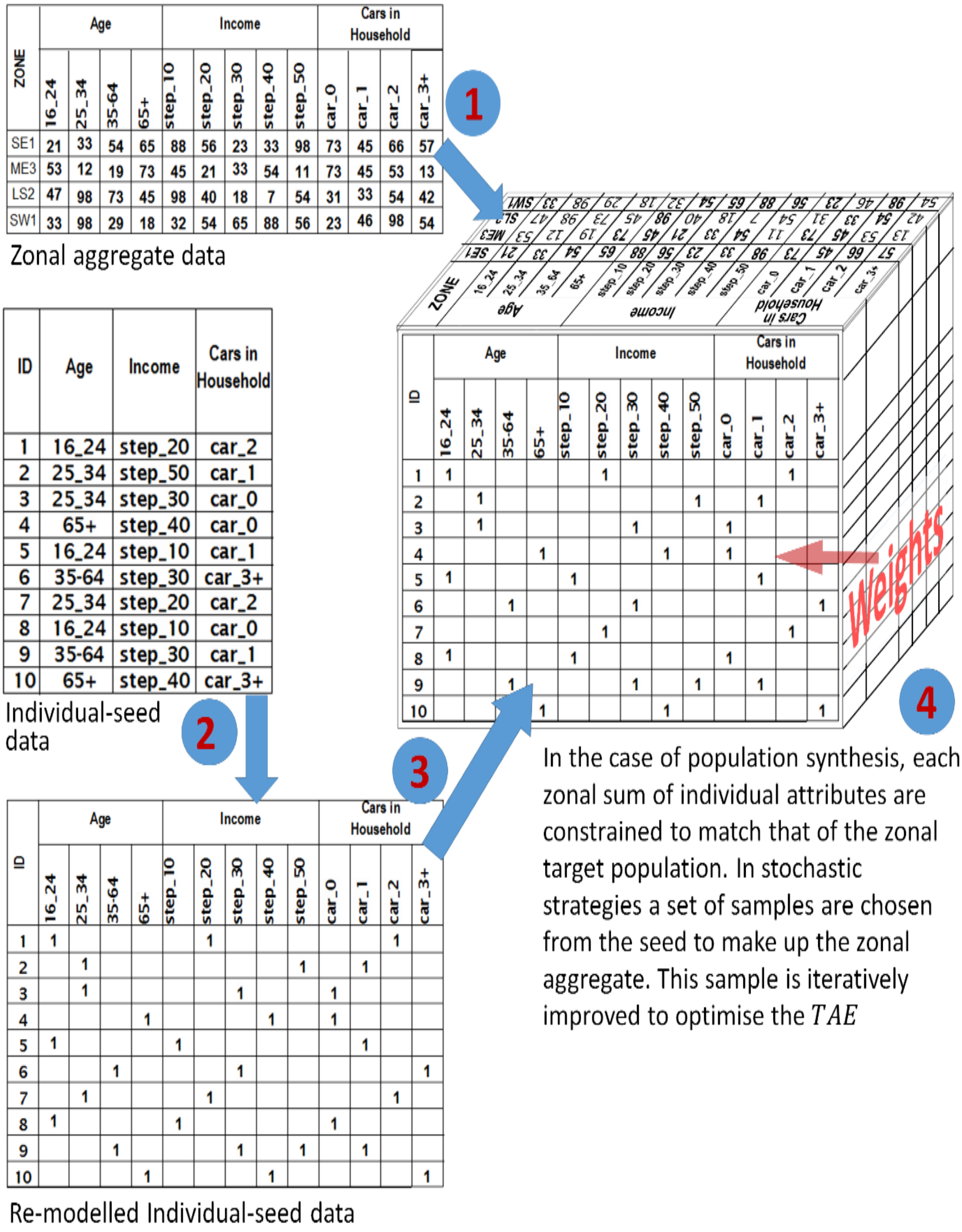
In practice, while the right individual-level table in Figure 4.2 represents the only structure permissible for the seed data, the left side represents the range of marginal data structures for the aggregate data. In essence, the constraints can be any combination of $1D$, $2D$, $3D$, up to nD where $(0 \leq n \leq \infty)$. This effectively implies that the constraint could be made up of several differently multi-dimensional arrays. These arrays are effectively partial views of the full joint distribution of the variables. The 2011 Census interaction data implies a marginal (Upton, 1985) constraint made up of eight (8) sets of $3D$ slabs, a slab for the age, gender, quasi-income, ethnicity, commute mode, household-type, -cars and children, cross-tabulated against the residence and destination variables and categories therein. Although the seed data is limited to the tabular depiction, a D-column table $(0 \leq D \leq \infty)$ is in fact a D-dimensional structured array.

The intuition behind creating a detailed micro-level population from an individual-level seed and zonal aggregates of the population, is that if a zone consists of an aggregate of say 20 people, with particular zonal characteristic, for instance that they are mostly aged and on high incomes: then if a seed sample representative of the entire region is available, the zone can be reconstituted from the sample by making an optimized selection of aged people on high incomes from the sample. To fulfil the constraint of sustaining the volume of people in the zone, the limited available aged and high income individuals in the sample might have to be replicated many times, hence the concept of weights which are indicative of how many times a type of individual is replicated to fulfil the synthetic population. In the case of mobility interaction, the weight assigned to an individual would be indicative of how representative the individual is of passengers on a particular residence-destination flow.

When the seed sample used in spatial microsimulation are skewed and not representative of the wider target population, the weight simulated for an individual in spatial microsimulation then also reflects how representatively that individual was captured in the sample seed. If for example the population consists of 20 people, 10 aged and 10 young and if the sample seed is made up of only 1 young individual and 10 aged, then the weights assigned to the young individual would be 10, whilst that assigned to the aged would be 1. Such a result would not be reflective of the young individual being ten-fold representative of the zone; instead it simply implies that the commensurate proportion of the young person is missing from the seed by a ten-fold factor. The weight as such is indicative of a combination of how representative an individual is of a zone, and the relative proportion and representativeness of that individual in the seed.

4.4.2 Procedures of the methods

The difference in the alternative spatial microsimulation procedures lies in the numerical algorithm or probabilistic method adopted for estimating or calculating the weights assigned to individuals in the seed data. Typically, the seed sample consists of a broad range of attributes for each individual forming a rich variable joint distribution which is limited only by being a sample, instead of the whole population. On the other hand, the aggregate data (like the Census) is anonymized for confidentiality and the zonal attributes are aggregated to form a cross tabulation. This cross-tabulation creates the challenge of relating individuals within say an age category, to the same individual within the gender, income and the other demographic attribute categories of the seed sample. The solution is to optimise the choice of individuals to minimize the difference between the marginal totals of the aggregate data and the synthesized population. In other words satisfying the condition that the micro-population created from the seed has margins adjusted and constrained to the aggregate margins. In such a case the residual between the aggregate and the synthetic population is minimized. Figure 4.3 is used to illustrate the similarities in the deterministic and stochastic procedures.



ZONE	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
SE1	21	33	54	65	88	56	23	33	98	73	45	66	57
ME3	53	12	19	73	45	21	33	54	11	73	45	53	13
LS2	47	98	73	45	98	40	18	7	54	31	33	54	42
SW1	33	98	29	18	32	54	65	88	56	23	46	98	54

Zonal aggregate data

ID	Age	Income	Cars in Household
1	16_24	step_20	car_2
2	25_34	step_50	car_1
3	25_34	step_30	car_0
4	65+	step_40	car_0
5	16_24	step_10	car_1
6	35-64	step_30	car_3+
7	25_34	step_20	car_2
8	16_24	step_10	car_0
9	35-64	step_30	car_1
10	65+	step_40	car_3+

Individual-seed data

ID	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
1	1					1						1	
2		1							1		1		
3		1					1			1			
4				1				1		1			
5	1				1						1		
6			1				1						1
7		1				1						1	
8	1				1					1			
9			1				1		1		1		
10				1				1					1

Re-modelled Individual-seed data

ID	Age				Income					Cars in Household			
	16_24	25_34	35-64	65+	step_10	step_20	step_30	step_40	step_50	car_0	car_1	car_2	car_3+
1	1					1						1	
2		1							1		1		
3		1					1			1			
4				1				1		1			
5	1				1						1		
6			1				1						1
7						1						1	
8	1				1					1			
9			1				1		1		1		
10				1				1					1

In the case of population synthesis, each zonal sum of individual attributes are constrained to match that of the zonal target population. In stochastic strategies a set of samples are chosen from the seed to make up the zonal aggregate. This sample is iteratively improved to optimise the TAE

Figure 4.3 Re-modelling in spatial microsimulation procedures.

An explanation of the IPF stages are detailed in the literature (Lovelace and Dumont, 2016). With reference to Figure 4.3, the seed data in deterministic methods have a structure typical of the middle left table in Figure 4.3. This seed data table is remodelled into a binary table (bottom left table) such that it has the same variables and variable categories as the target table (shown as the upper left table). The result is that the seed table is re-structured to produce the lower-left table in Figure 4.3. There are several software algorithms in literature that are suited to perform the restructuring (Dowle et al., 2017). As a result of the re-structuring, the individual-level seed and aggregate tables now have the same column headings, and this form the basis for reconciling the two tables (the seed and the target). The reconciliation yields the 3-D array on the right of Figure 4.3. Following this, considering one zone at a time, the sample seed is iteratively multiplied to fulfil the attributes of each variable category in each zone. This process is iteratively repeated for each variable category, within each variable in turn within each geographic zone until the results do not change substantially, yielding convergence. The criterion for convergence adopted is applied to the objective function, and is set up such that the difference between the distribution of the resultant simulated population and the target population (typically the Census) is minimal, and hence optimized.

The proposal distribution used in the deterministic procedure consists of the entire seed sample. This seed sample is simply replicated iteratively to fulfil target distributions for each category of variable and within each geographic zone. The deterministic procedure is conceptually different from the stochastic procedures in that in the deterministic process of replicating the passengers to fulfil the target distribution, fractions of individuals are used to achieve a more accurate solution. The mechanism of deterministic procedures is to specify a proposal population from all of the seed, and multiply these by weights to replicate the distribution of the target. The proposal distribution is iteratively improved subject to the objective function, until convergence to a target distribution. The ability to select fractional individuals ensures that non-absorbing regular iterative transitions are created (Bajram Spaseski and Ler, 2013) at each iteration step, ensuring convergence to the stationary target distribution. Widely used deterministic strategies are reported in literature (Barthelemy et al., 2016, Lovelace and Dumont, 2016).

The procedure for stochastic spatial microsimulation can also be illustrated using the data tables in Figure 4.3. In the Metropolis-Hastings based stochastic cases, i.e. hill climbing (HC) and simulated annealing (SA), once the seed is re-structured and the array on the right in Figure 4.3 is created, then, for each geographic zone, the seed is sampled forming the proposal distribution. The volume of the sample selected is equal to the population of the zone. This selection is then improved as specified by the particular stochastic algorithm adopted (i.e. HC or SA), until an optimum replication of the target population is achieved. The general procedure is as follows: if the marginal indicates a zone has a population N, the hill climbing method randomly selects an initial choice set of size N from the individual seed (creating the first proposal distribution), then the procedure randomly selects an element N_i from the proposal N to be consider for replacement, and an element 'D' from the sample seed as a potential replacement for N_i. If 'D' improves the choice set by improving the objective function i.e. by reducing the total absolute error[24] (TAE), then it is retained, otherwise it is discarded. The process outlined above is repeated until a convergence criterion is met. A known disadvantage of this HC version (Morris, 1993, Williamson et al., 1998) is that the process could converge irreversibly to a sub-optimal solution, a so called 'local optima'. This will result in a sub-optimal simulated population for a zone whereby a more radical change of the choice set might have further improved the accuracy (TAE) of the simulated population. Alternative HC algorithms exist whereby all the random selections are with replacement, and the cycle consists of repeatedly assessing the contribution of each element N_i of sample N to the TAE, and only the worst contributor is put up for potential change. The ever forward steps of the HC algorithm (at the detriment of getting stuck at a local minima) is what gives it the name hill climbing, with the classic potential entrapment at foothills if the climb does not incorporate back-stepping in the process of a search for the optimum (i.e. the top of the hill).

To circumvent the entrapment associated with the HC algorithm, the simulated annealing (SA) method which is akin to the Metropolis-Hastings MCMC is introduced, such that provision is made for forward and backward steps in the optimisation algorithm. The SA optimisation stage is mapped to

[24] The TAE which stands for Total Absolute Error, and is an evaluation criteria used in spatial microsimulation. The TAE is equal to the absolute value of the objective function at optimum. In cases where the object of the simulation is to minimize the absolute difference in simulated and target volumes, then the TAE is the sum of the absolute differences in volumes for all variable categories.

a law of thermodynamics given by $p(\delta E) = e^{-(\delta E/T)}$, such that T is defined as the maximum tolerable change in TAE. A measured change in TAE (i.e. δE) is accepted in the proviso that a randomly generated number exceeds the value calculated for $p(\delta E)$. The tolerance T is subsequently reduced after a defined number of changes in the initial seed have taken place, so the choice set (proposal density) gradually converges to an optimum. In practise, T can initially be set to a broad range of 1000, then an initial choice set is randomly selected, and the change in TAE (δE) assessed. Any change yielding an improvement in TAE is adopted, but a change worsening the TAE are further assessed by comparing the value of $p(\delta E)$ with a randomly generated number between 0 and 1. If $p(\delta E)$ is larger, then the worsening replacement element is adopted, otherwise rejected and the process cycle repeated until convergence.

The third generally use stochastic spatial microsimulation strategy is called the genetic algorithm (GA) which mimics the perceived process of evolutionary change in nature. The advantage of the genetic algorithms are that they ensure that the spatial microsimulation constrained optimisation solution does not get stuck at a local optima, stationary or inflection points as the entire solution space is repeatedly searched. The GA algorithm is not explored any further in this research because, it is a variant of the HC and SA methods, albeit including more parameters which are difficult to tune (Williamson et al., 1998). Further GA algorithms in practical problems do not yield better results than the HC and SA methods (Williamson et al., 1998).

4.4.3 Synopsis of methodology

A summary of the range of deterministic and stochastic spatial microsimulation methods are shown in Figure 4.4. These methods are explored in a controlled experiment to ascertain the behaviour of deterministic and stochastic spatial microsimulation methods in scenarios where the seed sample is biased or skewed, as would be the case with typical novel consumer and observational data. Recall that these datasets are skewed because they represent revealed activity from a sub-set of the population – only those that consume the particular digital service. Further studies on deterministic and stochastic spatial microsimulation methods investigate the effect of sample ratios and number of constraint variables, on the accuracy of spatial microsimulation results. This is a precursor for application of the methodologies to actual realistic data in the case study of railway mobility in the West Yorkshire study area.

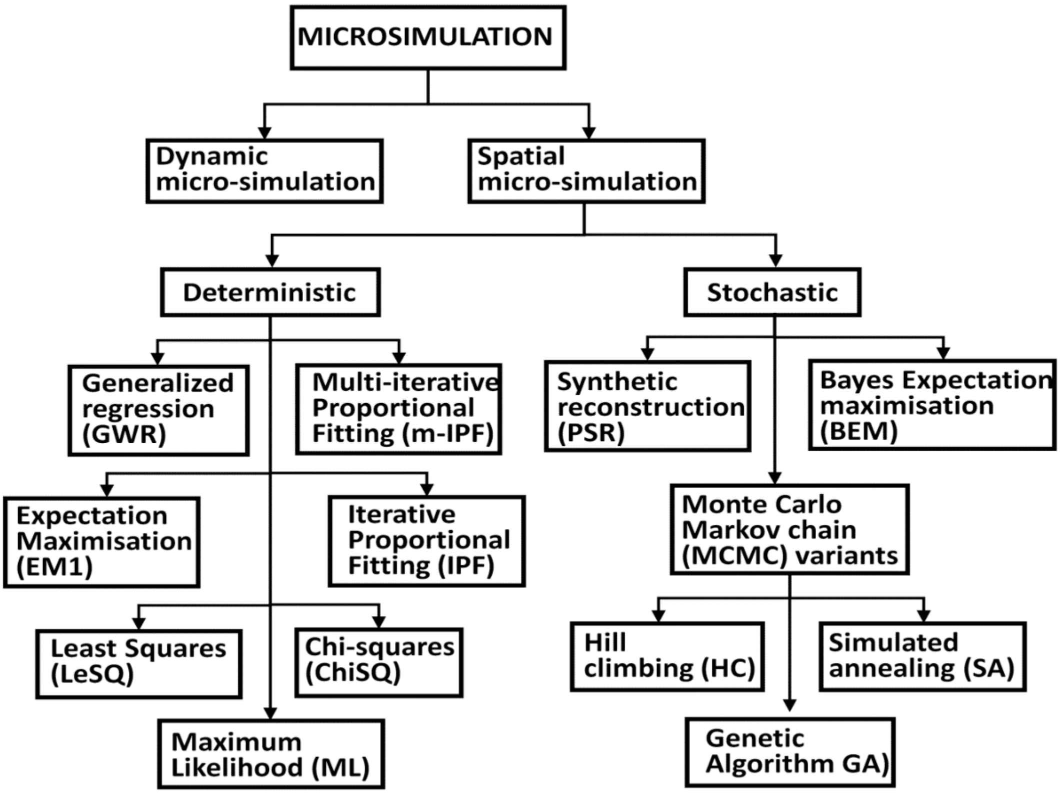

MICROSIMULATION
Dynamic micro-simulation
Spatial micro-simulation
Deterministic
Stochastic
Generalized regression (GWR)
Multi-iterative Proportional Fitting (m-IPF)
Synthetic reconstruction (PSR)
Bayes Expectation maximisation (BEM)
Expectation Maximisation (EM1)
Iterative Proportional Fitting (IPF)
Monte Carlo Markov chain (MCMC) variants
Least Squares (LeSQ)
Chi-squares (ChiSQ)
Hill climbing (HC)
Simulated annealing (SA)
Maximum Likelihood (ML)
Genetic Algorithm GA)

Figure 4.4 Range of spatial microsimulation methodologies.

4.5 Experiment on microsimulation methods

The previous sections have detailed a number of different microsimulation methods. Each have numerous advantages and disadvantages (such as the precision of deterministic methods, the likelihood to find local sub-optima in stochastic methods, and their limited computational power requirements). It is extremely difficult to determine which spatial microsimulation method would be the most appropriate for harnessing the relatively novel consumer data from a purely theoretical standpoint. Therefore controlled experiments will be conducted to assess the behaviour of deterministic and stochastic spatial microsimulation strategies. These experiments are conducted using only the NRTS[25] data for the West Yorkshire County. For the experiment, the 8 variables which have relevance to railway mobility (and are typical of those available in the Census interaction data) are selected from the NRTS. These

[25] National Rail Travel Survey (NRTS) was conducted in 2004/2005 in the UK, in areas outside London. The NRTS measured passenger characteristics like access and egress modes to railway network, passenger final destination and home address, purpose of travel, ticket type, and time of arrival at station, travel start time, as well as a range of socio-demographic characters.

variables are ‘age’, ‘gender’, ‘income’, ‘commute mode’, ‘household-type’, ‘household-cars’, ‘ethnicity’, and ‘household children’. From these variables, aggregate data arrays of different dimensions ($1D$, $2D$, $3D$ up to $8D$) are created, these represent the ‘true’ population (i.e. the target constraints as seen in Figure 4.2, Section 4.4.1). The seed on the other hand is created by taking tuple (row) samples from the NRTS data. A seed of low sample ratio is created by taking fewer samples from the NRTS dataset, a skewed seed is created by systematically taking samples from an ordered NRTS dataset, etc. To assess the propensity of a particular methodology to predict non-constrained variables for example, a 3D target could be used to generate a synthetic NRTS population from a random NRTS seed. This synthetic population is then compared with the original NRTS. The difference between the synthetic and actual population is assessed for those variables not included as part of the 3D target constraint. A Monte Carlo experiment is designed to test a range of deterministic and stochastic spatial microsimulation methods, for different configurations of the target and seed, enabling the sensitivity and accuracy of the spatial microsimulation methods to be assessed. A summary of the experiments consist of the following, with specific details included in the subsequent sections:

1. The number of target constraints are increased from 1 through to 7 (which is one less the full set of variables), and the effect of the number of constraints on the TAE[26] are assessed.
2. The predictive accuracy of the deterministic and stochastic methods are assessed for those variables not included as target constraints in the spatial microsimulation process.
3. Sample ratios of the seed are varied from 0.05%to 85%, for a target constraint created by aggregating a full set of the 7 variables.
4. The skew in the seed samples are progressively increased, while the resulting TAE values are compared for the deterministic and stochastic methods. The skew is achieved by taking a random sample and then various degrees of systematic samples, as described below.

The methodology of the experiment is presented below, detailing some the data preparation and procedures for the experiment.

[26] The total absolute error (TAE) is a measure of the fit between the simulated population and the actual target population. The absolute difference in count of individuals in each variable category for the synthetic and actual target populations are summed to yield the TAE.

4.5.1 Methodology of experiment

The experiment assessed the efficacy of deterministic and stochastic spatial microsimulation methods. The deterministic method tested is the predominant m-IPF (Barthelemy and Suesse, 2016). Alternative deterministic methods are the Chi squares, maximum likelihood and least squares methods (Barthelemy and Suesse, 2016, Deming and Stephan, 1940), however the implementations of these are based on the Lagrange multiplier method, and were found to be intractable for large datasets (>20,000 tuples) as used in the experiment, so these results have not been presented. The stochastic methods assessed are the simulated-annealing (SA) and hill-climbing (HC) methods which are respectively variants of the Metropolis-Hastings and Rejection sampling methods. The results presented are for the SA method, as the HC yields sporadic sub-optimal results, thereby precluding an ability to control the experiment. Recent simulation methods based on the Gibbs sampler have been developed (Farooq and Bierlaire, 2017), but discussion on these more recent stochastic methods are deferred to the next chapter, as they are not considered established enough to warrant comparison with the m-IPF, HC, and SA methods.

4.5.1.1 Settings and data

A preliminary summary of the NRTS variables chosen for use in assessing the deterministic and stochastic spatial microsimulation methods are presented in Figure 4.5. The summary is presented to highlight that a range of data variables are normally distributed, albeit being skewed relative to the wider population (as the NRTS consists of a larger percentage of commuters relative to Census). Different sampling regimes, i.e. systematic and random are adopted to enable the data to be tailored for the desired experiments.

The plots shown in Figure 4.5 are for NRTS flows emanating and terminating in the county of West Yorkshire. Going from left to right on the upper row, and similarly from left to right on the lower-row, a summary of the data show the following: Income has been divided into seven brackets (step-10 to step-70[27]), as specified in the NRTS questionnaire (DfT, 2013a). The display of volumes of passengers in each socio-demographic category across the

[27] The income values are based on values in the period 2001 and 2004/2005. Income bracket step-10 earn lower than £7000, step-20 earn £7000-£12,500, step-30 earn £12,501-£17,500, step-40 earn £17,501-£35,000, step-5 earn £35,001-£50,000, step-60 earn £50001-£75,000, and step-70 earn more than £75,000.

ordinal income bands show normal distributions with peak volumes associated with the step-40 income bracket. Expectedly the income-income plot indicates a mode income in the step-40 category (£17,501-£35,000 in 2004). As the NRTS survey was conducted on the railways, a minority of the passengers captured do not commute by train (NTC), with a higher volume of passengers commuting by train (TC) of about 66%. This distribution is quite unlike the wider population which reveals a lower percentage (~8%) of passengers commuting by rail (ONS, 2013). The age-income relationship shows a positive increasing trend, apart from the over 65year age bracket where the income reverts to the lower income categories. Both the male and female genders are normally distributed about the modal step-40 income category. The slightly higher volume of female passengers is indicative that more males use other means (perhaps the car) for mobility and commute. Car ownership increases with income, and income increases as household type goes from A to D, indicative that more mature households tend to have more cars available at the disposal of household members. The 16-24yrs bracket indicates a fair number with over 3 cars, and this is perhaps indicative of young adults in established households where there are 3 cars or more. Expectedly, the majority of the passengers are of European descent

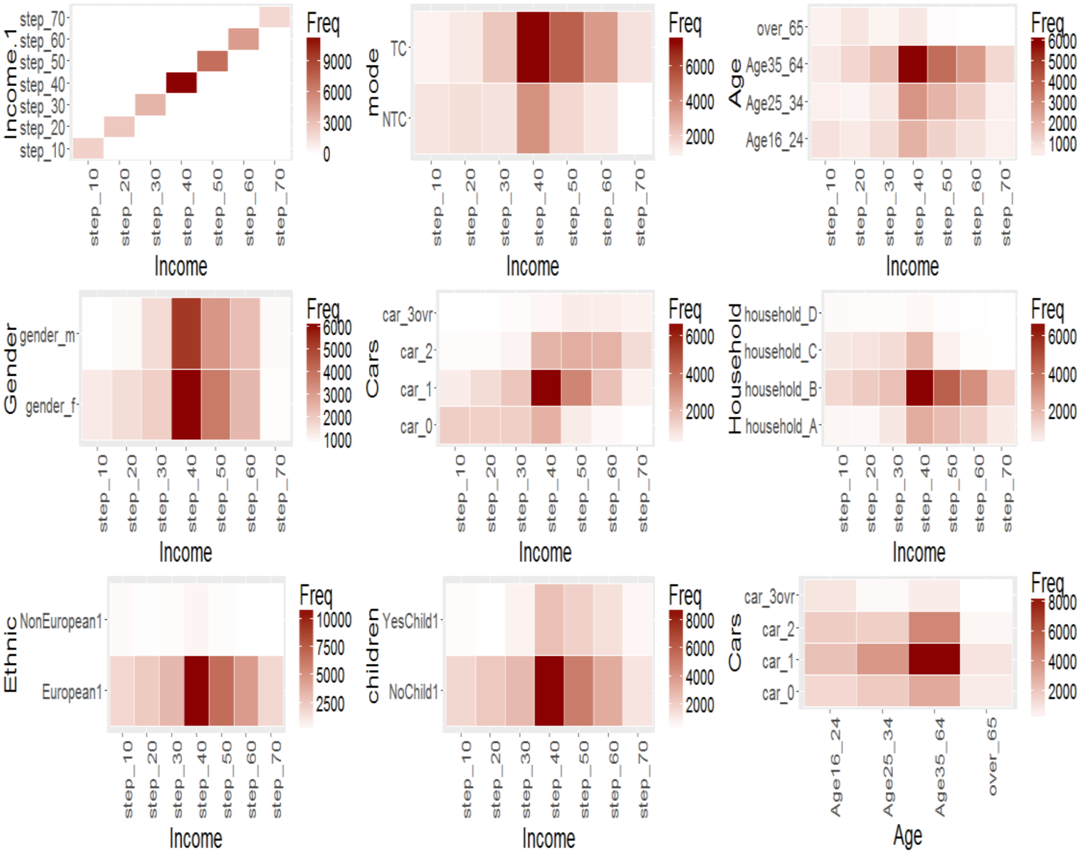


Figure 4.5 Cross-table of categorical NTRS covariates.

4.5.1.2 Data preparation

For the purposes of managing the computational demands of the spatial microsimulation algorithms, the NRTS data which was originally zoned to Postcode Sector boundaries were spatially aggregated to Postcode Areas. It is assumed that this did not affect the results of the experiment as both the deterministic and stochastic methods are equally affected. Postcode boundaries in the UK start from the larger Areas (about 120 of them covering the UK), to Districts, Sectors (see Figure 3.6 for West Yorkshire), and then smaller Postcode Units (containing about 15 addresses). For the spatial microsimulation experiment, the Postcode Sectors in West Yorkshire which originally numbered about 350 were spatially aggregated to the next lower geography Areas. This resulted in the 5 Postcode Areas shown in Figure 4.6. The NRTS seed samples are about 65000 tuples for the West Yorkshire population of about 2.4m (a sampling ratio of about 3%). As such, small Postcode Sectors like LS52 and BD127 with populations of about 1,600 and 4,300 respectively, only have 3 NRTS samples each. The low samples (only 3) were not deemed representative of either of the 1,600 or 4,300 zonal populace; hence the further need to aggregate the volumes, facilitating adequate representation of a geographic zone. The resulting Postcode Areas (BD, HD, HX, LS, WF) covering West Yorkshire have NRTS samples ranging from of 2,599 (in HX) to 24,121 (in LS) with a variance of 6.5E+07.

To restrict the analysis to within the West Yorkshire county, Postcode Areas were clipped, and as a consequence there were nominal remnants of some peripheral Postcode Areas, and these have been reflected in the legend shown in Figure 4.6. For instance, in the left hand corner of the West Yorkshire county, the map when displayed using a GIS indicated that parts of the Oldham Postcode Area (OL148, OL138, OL145 etc.) were captured and were included in the analysis. Similarly, other sub-sections of other Postcode Areas were captured, including Sheffield, York, Harrogate, Blackburn and Doncaster (S, YO, HG, BB and DN). Perhaps these might have been removed from the analysis as these sub-sections are not representative of the Postcode Areas referenced. They have however been included in the analysis to replicate situations where a zone has a low number of counts compared to other typical values in the analysis. In effect, inclusion of these marginal Postcode Areas enables an assessment of the sensitivity of the different spatial microsimulation strategies to nominally low sample counts from a particular geographic zone.

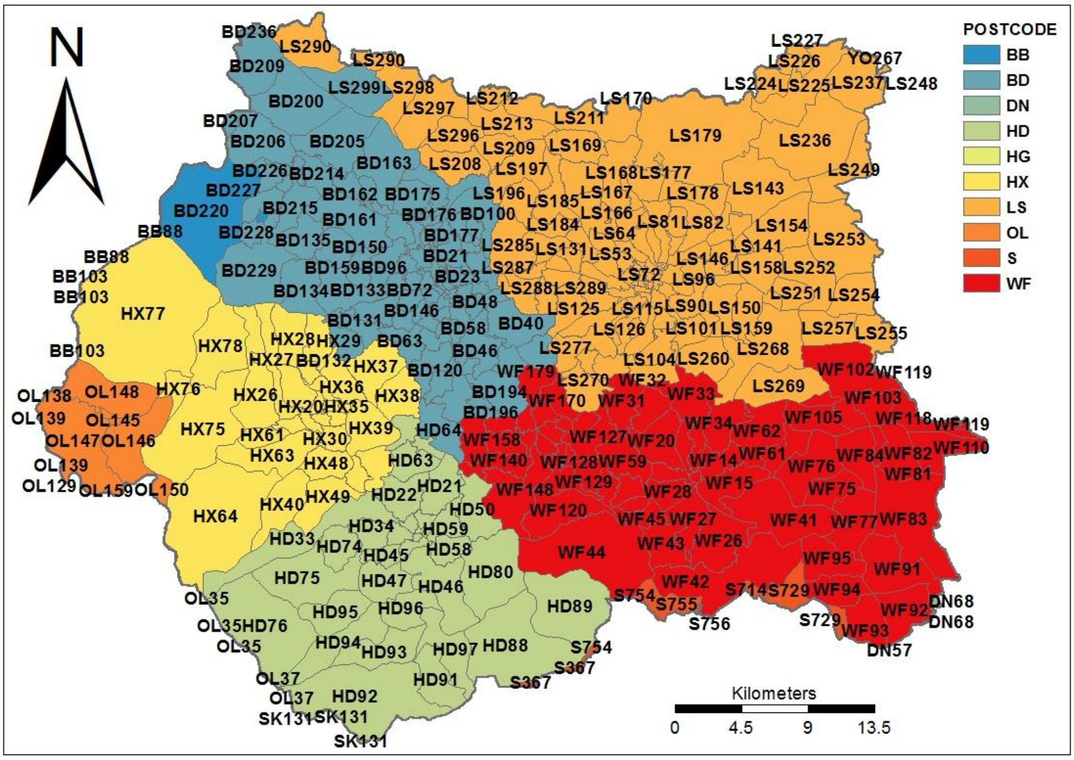


Figure 4.6 Re-zonation of Sectors to Areas (Postcodes).

4.5.1.3 Experiment procedure

The procedure implemented to assess various spatial microsimulation strategies is illustrated in the flowchart in Figure 4.7. The state-of-the-art deterministic m-IPF (Barthelemy et al., 2016) and established stochastic SA and HC methods (Kavroudakis, 2015) were assessed for suitability for use under various scenarios. The scenarios are first, when sample ratios of the seed are varied from 0.05% to 85%, second when the number of target constraints are increased from 1 through to 7, third when the skew in the seed samples are progressively increased, and fourth the predictive accuracy of the deterministic and stochastic methods are assessed for those variables not included as target constraints in the spatial microsimulation process. In each case, the total absolute error (TAE) is measured in a Monte Carlo experiment. The Monte Carlo experiment was necessary to ensure that the results revealed were not a feature of a particular choice of seed or target data. For example, in the experiment to assess the effect of variations in sample ratio, when a 55% sample is required to be chosen, the experiment is repeated severally[28] (350 times) for different 55% samples,

[28] 350 repeats were adopted in the software code developed. However, a sensitivity analysis indicated that there were no noticeable changes in the results over 100 repeats.

with randomly different seed values. This ensures that the distribution of TAE values resulting is a true reflection of the range of influences that a 55% sample could have on the TAE for deterministic and stochastic spatial microsimulation methods.

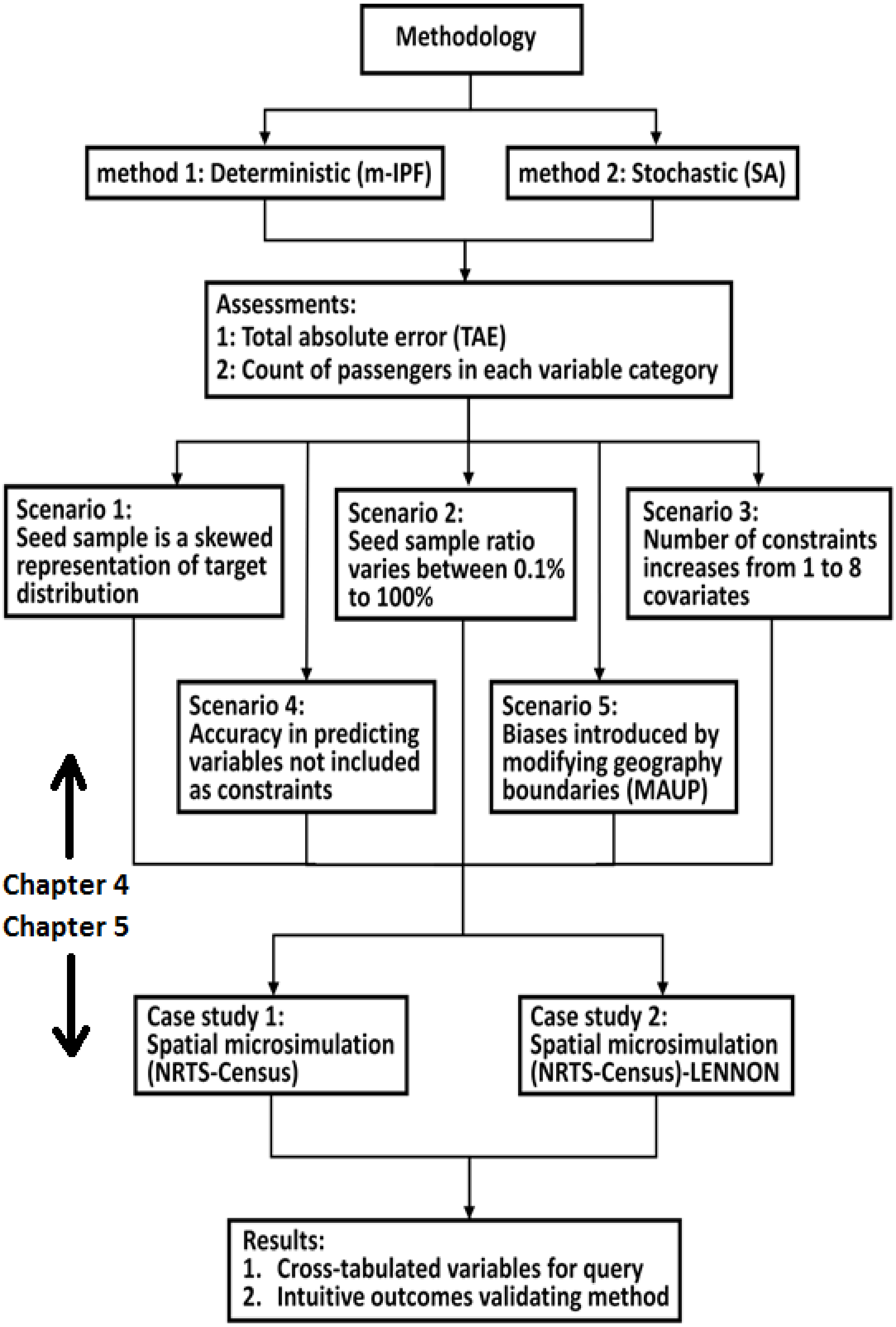

Methodology
method 1: Deterministic (m-IPF)
method 2: Stochastic (SA)
Assessments:
1: Total absolute error (TAE)
2: Count of passengers in each variable category
Scenario 1:
Seed sample is a skewed representation of target distribution
Scenario 2:
Seed sample ratio varies between 0.1% to 100%
Scenario 3:
Number of constraints increases from 1 to 8 covariates
Scenario 4:
Accuracy in predicting variables not included as constraints
Scenario 5:
Biases introduced by modifying geography boundaries (MAUP)
Chapter 4
Chapter 5
Case study 1:
Spatial microsimulation (NRTS-Census)
Case study 2:
Spatial microsimulation (NRTS-Census)-LENNON
Results:
1. Cross-tabulated variables for query
2. Intuitive outcomes validating method

Figure 4.7 Investigation of spatial microsimulation methods.

The implementation of the deterministic IPF procedure involves optimizing the fitting between a sample seed and a target population. The target seed would typically come from a sample taken from the wider population. The target would typically be derived from aggregate tables within the Census. During the standard IPF procedure however, the seed is actually fit to the marginal of the target table. As such IPF procedures in their standard form do not exploit the full joint (conditional) distribution of the target, but rather just the marginal distribution. In interpreting the accuracy of the results derived, these considerations need to be borne in mind. To date, there are no practical implementations of the IPF that have exploited the full conditional distribution of the target, and these form the crux of the methodology developed in the next chapter. The stochastic spatial microsimulation procedures on the other hand, inherently exploit the full joint distribution of the target, as samples are drawn from the seed and the count of all variable categories directly compared with the target. These considerations are discussed further at the conclusion of the range of Monte Carlo experiments reported in this chapter..

4.5.2 Results on deterministic and stochastic methods

The results presented are the distribution density and heat maps of the TAE values when a particular experiment is repeated multiple times in a Monte Carlo experiment. Regions of high TAE density are reflective of the predominant modal TAE value. The TAE density plot also reflects the sensitivity to particular changes for each deterministic and stochastic spatial microsimulation method. The R-Software (RStudioTeam, 2016) developed to implement each of the experiments have been discussed in Appendices A.1 to Appendix A.5 (and included in the CD accompanying the thesis).

4.5.2.1 Effect of increasing the number of constraints

There are $8P_3$ ways of choosing a set of options of three (3) from eight (8) variables (assuming order is also important). To establish the effect of the number of constraint variables, a Monte Carlo sampling was implemented on the choice set of constraints and the distribution of the TAE's for each sample set displayed in the density plots in Figure 4.8 and Figure 4.9. 100 different combinations of 1-way, 2-way, 3-way, up to 6-way array tables were separately created to form the target, and each of these implemented in the spatial microsimulation. Initially, the same 5% random sample of the NRTS was created forming the seed used for all the deterministic and stochastic spatial micro-simulation exercises. However, a sensitivity analysis showed that there were no marked changes in the trend of the results if the sample ratio (η) of the seed was increased above 5% $(5\% \leq \eta \leq 85\%)$. Similarly, there were no changes in the trend of results if several combinations $((\xi))$ of the 5% random sample $(1 \leq \xi \leq 100)$ was adopted as the seed, and the results presented as a 3D map. This Monte Carlo experiment as such smoothed out the effect of a particular choice set of variables and/or samples, thereby objectively revealing the effects of increasing the number of target variables for the deterministic and stochastic spatial microsimulation methodologies.

The results show in Figure 4.8 and Figure 4.9 reveal that for the m-IPF and SA methods, as the number of constraints increase, the TAE values are seen to reduce (for both the m-IPF and SA methods) as indicated by a drift of the density plots towards the origin (to the left) as constraints increase. The $TAE's$ for the stochastic method (SA) are considerably lower than those for the deterministic m-IPF. As observed in Figure 4.8 the results for 1 target constraint (**1-Constr**) are not reported on the plots due to the instability in the IPF method when only one constraint variable is used. It was found that for particular choices of 1-way $(1D)$ constraints, and for low sample ratios ($<$ 5%), the deterministic IPF had a software crash, due to the seed being complete non-representative of the particular choice of target, resulting in a simulated population of zero.

The large TAE values for the m-IPF are perhaps due to two issues. First, the deterministic methods only exploit the marginal of the target constraint, whilst the stochastic methods exploit the full joint conditional distribution of the target constraint. The use of marginal in the deterministic method effectively reduces the constraint requirement on the TAE results, thereby making provision for a more flexible sub-optimal range of results. This has been an issue reported in the literature concerning the IPF methodology (Farooq and Bierlaire, 2017). In Chapter 5, a methodology is developed for overcoming this particular short-coming of the IPF.

The second reason for higher TAE values from the m-IPF relate to the sensitivity of IPF's to small sample ratios due to the inclusion of peripheral zones (YO, S, OL) which have fewer and poorly representative samples. For each zone (Postcode Sector) in the geography, the SA method draws samples from the whole NRTS of size equal to the population of the zone being simulated (the proposal distribution). This sample is then iteratively improved by taking further individual samples from the NRTS, to replace elements of the proposal distribution, such that the proposal distribution is iteratively improved. The same result would be achieved more efficiently if the initial proposal distribution for a zone were formed by only those samples that had pertinence or relevance to the zone being simulated. As such, even if the initial sample for the SA were chosen from the entire NRTS, subsequent SA iterations to improve the sample yielding the simulated population, would narrow this sample down to those relevant to the particular zone being simulated. The m-IPF start with a sample (proposal) made up of the entire NRTS. Subsequent iterations of the m-IPF map this sample onto the marginal distribution of the population of the zone being simulated. As a result elements of the whole NRTS that are not relevant to a zone are filtered out in the mapping process as they would have no corresponding values in the target to map onto.

As a result of this mapping, the m-IPF would tend to result in fewer simulated elements, corresponding to only those that were able to be mapped onto the target population. As such, for the m-IPF, if a particular constraint cell lacked a representative in the sample, then IPF would not converge to a full population, whereas the stochastic SA method satisfices[29]. The stochastic methods would generate a full population for a zone once the aggregate targets are defined for the zone. This is the case even when the zone was poorly[30] sampled. In a commensurate m-IPF strategy on the other hand, the full volume of the zone would not be simulated, thereby yielding particularly high (poorer) TAE values.

[29] The concept of convergence is relative. When methods like the SA satisfices for instance, it creates a full but sub-optimal population. This is the case because the method starts off with a proposal distribution that is made up of the full requisite population volume. The SA methodology simply iteratively improves this proposal by replacing the elements of the distribution, one at a time, on each occasion maintaining the requisite volume of the zone. In essence, an argument could be made to the effect that full SA convergence had not been achieved as the simulated population does not exactly match the target population. The TAE measure is as such devied to describe the extent of the convergence. On the same basis, the m-IPF that did not yield the requisite population can not strictly be described as not having converged, as the methodology had converged as best it could, yielding the higher TAE value. In many of these deterministic and stochastic procedures, convergence is adduced after a pre-defined number of iterations.

[30] Typically a zone would have at least one sample, and with the stochastic methods such a sample would bear a high weight to create the full simulated population. In a commensurate m-IPF, the created weight will be lower, as the seed is replicated to yield just the requisite volume of similar individuals in the zonal population.

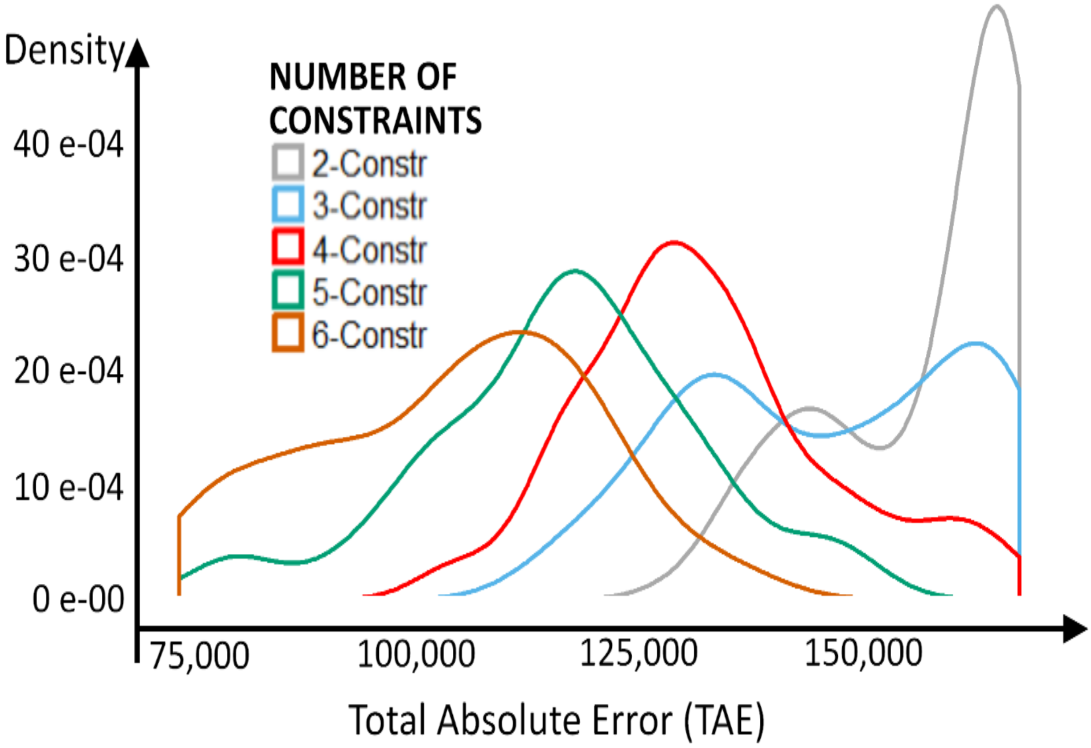


Figure 4.8 TAE vs No. of constraints (deterministic m-IPF)

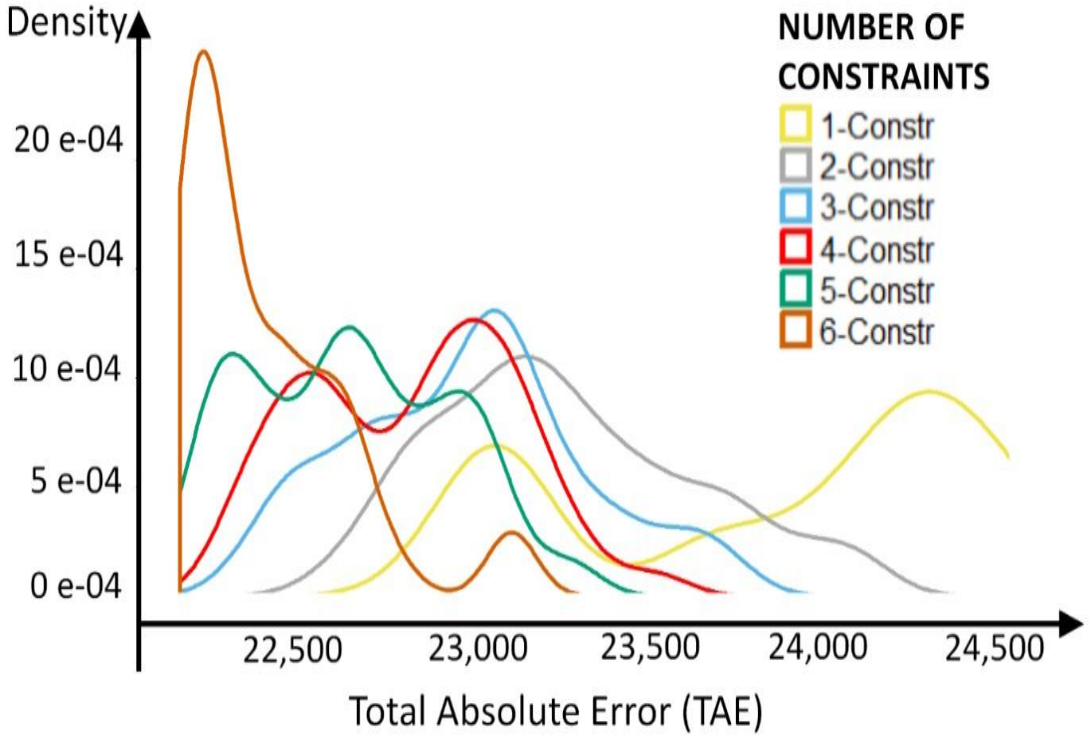


Figure 4.9 TAE vs No. of constraints (stochastic SA).

4.5.2.2 Sensitivity of TAE to population of a zone

The mechanism through which the TAE is generated is explained using the heat maps presented in Figure 4.10 and Figure 4.11, respectively for TAE results generated from the deterministic m-IPF and the stochastic SA methods. The heat map on left of Figure 4.10 for the deterministic method shows the variation of TAE for each simulated zone, as the zonal population decreases from the left to right of the horizontal axis. The heat map on the right (of Figure 4.10) is a similar measure of the TAE, but this time the TAE is normalised by the zonal population. The heat maps in Figure 4.11 are corresponding TAE values generated from the stochastic SA procedure. The left being the normal TAE, whilst the right represents a normalised TAE value.

In both the deterministic and stochastic procedures, as seen from the plots on the right, the regions of higher normalised TAE values are those zones with small populations, and this effect is commensurately more pronounced in the results for the deterministic procedure (the right of Figure 4.10). The default results are shown by the plots on the left, whereby it is observed that zones with higher population tend to yield commensurately higher TAE values. This effect is again more pronounced for the deterministic m-IPF method as shown on the left of Figure 4.10. Whilst the higher population zones have higher total absolute errors values ($TAE's$), when the TAE is normalised by the population of the zone, the trend is reversed. This indicates that lower population zones are more susceptible to spatial microsimulation errors, and for practical purposes, this suggests that a larger sample size renders a lower susceptibility to spatial microsimulation errors. Since the deterministic m-IPF method is more sensitive to this effect, it yields higher TAE values as reported earlier in Figure 4.8 and Figure 4.9.

A further issue observed from the heat maps in Figure 4.10 and Figure 4.11 reveal that variables with fewer categories record higher error consistency (and lower variability) across the variables, as seen for instance in the commute mode (TC/NTC), ethnicity (Europe/Non) and household children (Yes Child/No Child) variables. The variables with many categories are seen to record richer variability as seen for instance in the household-type or the cars variable. This is additionally indicative that the richer a variable in terms of its number of categories, the better its ability in extracting the heterogeneity in a simulated population.

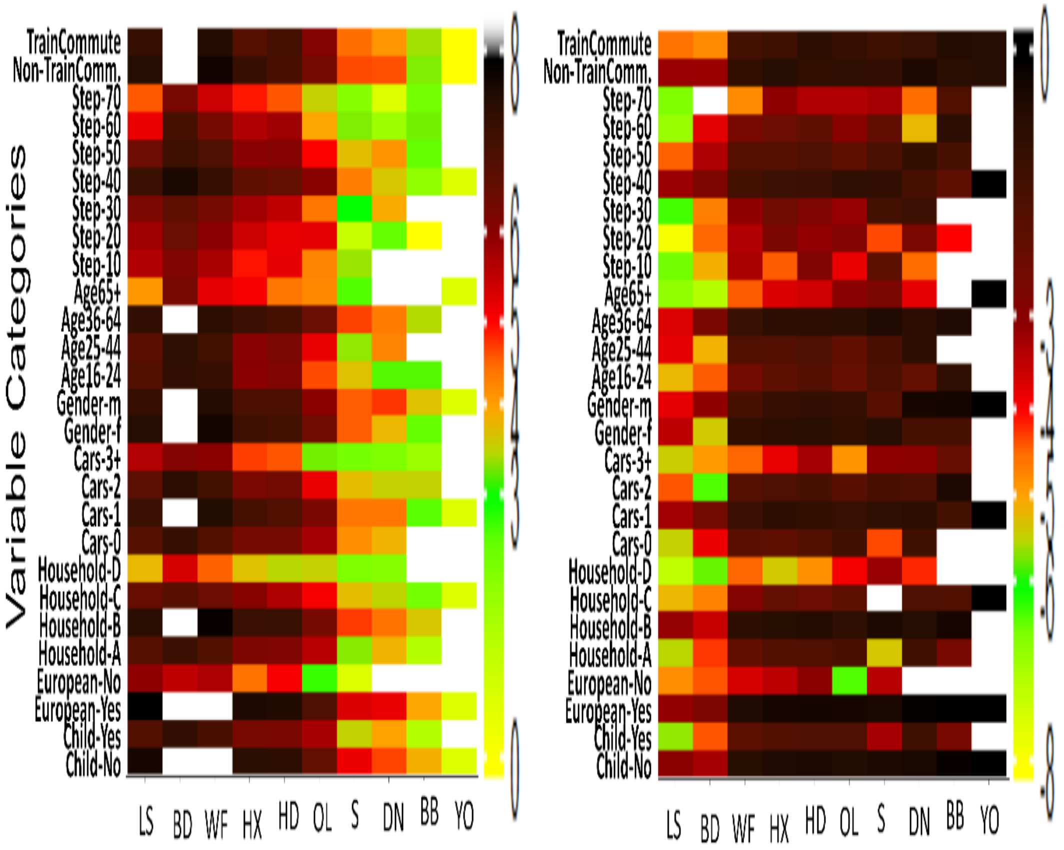


Figure 4.10 TAE for decreasing zonal population (L-R) - deterministic.

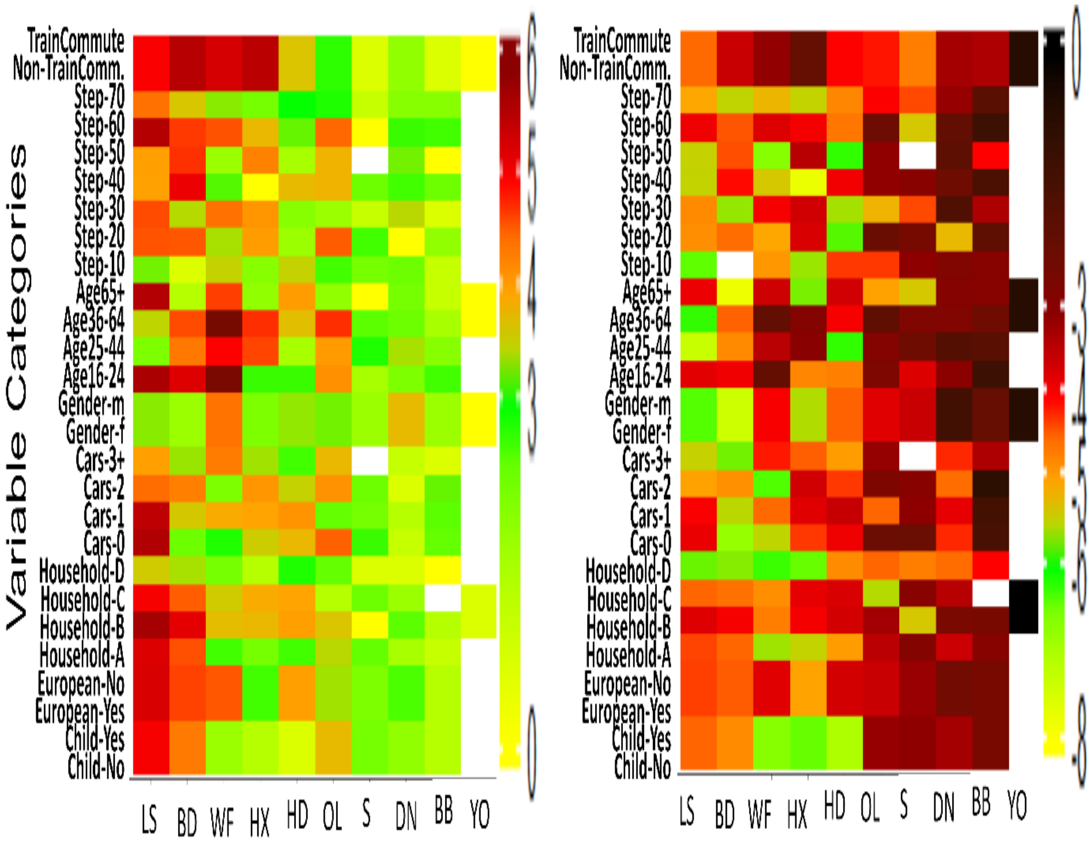


Figure 4.11 TAE for decreasing zonal population (L-R) - stochastic.

4.5.2.3 Prediction of TAE for non-constrained attributes

The second comparison of the deterministic m-IPF and stochastic SA methods addresses the question: how well predicted are the values of those variables not included as constraints in the spatial microsimulation? In essence, if subsets of variables within the NRTS are used as target constraints during the spatial microsimulation, how well are the simulated zonal aggregates predicted for those variables that were not included in the set of aggregate target constraints? In essence for example, if the age, income, gender and commute-mode variables are used to create the aggregate cross-table constraints, how well would the different spatial microsimulation methods predict the other variables household-type, cars, ethnicity, and household-children, as they were not used to create the aggregate constraint? This question resolves the issue of the robustness of the different spatial microsimulation strategies since the synthetic population traditionally inherits the attributes of all the datasets involved, irrespective of if those variables that constituted the constraint.

The procedure for predicting the TAE values for those variables not included as constraints in the SA and m-IPF methods are as follows: For the example case where the aggregate target constraint is created from only the age, income, gender and commute-mode variables. This would yield a 4-way ($4D$) aggregate array. The deterministic m-IPF and stochastic SA spatial microsimulation methods are then implemented to create the two sets of simulated populations for the geographic zones (one set from m-IPF and another from SA). Note that the identity information are known for all the individuals forming the simulated population, as this detail was carried along in the simulation process. For each zone, the individuals simulated are aggregated by variable categories. For instance for Zone 1, using the m-IPF method, there could be 34 individuals simulated in the age category Ag16-24, and 27 people simulated in the age category Ag25-34. Using the SA method these simulated individuals might amount to 46 in the Ag16-24 age band and 35 in the Ag25-34 band. The actual NRTS population might have indicated that for Zone 1, there are only 42 people in age category Ag16-24 and 33 people in the Ag25-34 band. In such an instance the total absolute difference (TAE) for the populations created using the m-IPF and SA methods, would be 14 and 6 respectively. It is noteworthy to point out however, that as the identities of the individuals simulated using the m-IPF and SA are known. As such, the remaining attributes of these individuals can easily be read from the original NRTS table.

As a result, the volumes of these individuals within the variable categories household-type, cars, ethnicity, and household-children (not used as constraints in the spatial microsimulation) can be deduced. For each zone, the resulting volumes simulated for each of the unconstrained variable categories is presented in tabular form (creating a table of zones versus variable categories). The TAE is estimated as the absolute difference in the two tables created from m-IPF and SA. These result in the TAE values predicted for the unconstrained variables using the m-IPF and SA methods. Using this method, TAE values are deduced as the number of constraints increase from a 1-way $(1D)$ aggregate to a 7-way $(7D)$ aggregate, increasingly using the ordered sequence of variables (1) household-children, (2) ethnicity, (3) household-type and (4) cars, (5) gender, (6) age, (7) income, and (8) commute-mode. When only a 1-way constraint is adopted, that would be household-children, when a 2-way constraint is used that would be household-children and ethnicity. A 3-way constraint would include the household-type variable, forming a 3-way (3D) constraint made up of household-children, ethnicity, and household-type. The same logic follows for an increasing number of constraints.

When only the 1-way constraint is used, the TAE is calculated separately for the variable categories in that particular constraint (i.e. NoChild, and YesChild). The TAE is also calculated for the variable categories of all the other variables not included as the constraint and these variables would be (2) ethnicity, (3) household-type and (4) cars, (5) gender, (6) age, (7) income, and (8) commute-mode, with respective variable categories (Europe and Non-Europe, Household-A, Household-B, Household-C, and Household-D, Car-0, Car-1, and Car-2, and Car-3+, Female and Male, Age16-24, Age25-34, Age35-64, Age65+, Income-10, Income-20, Income-30, Income-40, Income-50, Income-60, and Income-70, No-train, and Train). For say a 5-way constraint, the first five variables are used as the constraint i.e. (1) household-children, (2) ethnicity, (3) household-type and (4) cars, and (5) gender. The TAE is reported for the variable categories within these first five variables (i.e. (1) household-children, (2) ethnicity, (3) household-type and (4) cars, (5) gender), as well as the variable categories for those variables not included in the constraint set (i.e. (6) age, (7) income, and (8) commute-mode) with categories Age16-24, Age25-34, Age35-64, Age65+, Income-10, Income-20, Income-30, Income-40, Income-50, Income-60, and Income-70, No-train, and Train. These results are visually summarised using the stacked plots presented in Figure 4.12 and Figure 4.13.

The stacked column charts In Figure 4.12 and Figure 4.13 show the TAE values for the different numbers of constraints. The categories of the constraint variables are listed along the horizontal axis. 1 constraint refers to when the first variables (household-child) and their categories (No-Child and Yes-Child) are used to build the aggregate constraint. 3 constraints refer to when the first three variables (household-child, ethnicity, and household-type) and their categories (No-Child and Yes-Child, Europe and Non-Europe, and Household-A, Household-B, Household-C, and Household-D) are used to build the aggregate constraint. This concept is similarly applied for 5, and 7 constraints. When 1 constraint is applied (the pink-orange parts of the stacked columns), it is seen that whilst the deterministic m-IPF procedure yields no particular trend across the range of variable categories, the stochastic SA yields lower TAE values for the constraint variables (No-Child and Yes-Child), with increasing values of TAE for the other variables. Similarly for 3 constraint variables (household-child, ethnicity, and household-type), shown by the green parts of the stacked columns, the deterministic m-IPF does not show any particular trend in TAE values across the variables categories on the horizontal axis. The stochastic SA method however, shows that the TAE values are lower for the 3 constrained variables (household-child, ethnicity, and household-type) as seen by the green parts of the stacked columns for the categories No-Child and Yes-Child, Europe and Non-Europe, and Household-A, Household-B, Household-C, Household-D. This scenario of a lack of trends in the deterministic TAE values, and the trend in the stochastic TAE values can be seen to continue when there are 5, and 7 variable constraints.

The results in Figure 4.12 is indicative that the predictive results for variables not included as constraints in the deterministic m-IPF procedure, are less sensitive to the particular choice set of variables. The stochastic SA method on the other hand, reveals relatively lower TAE values for those variables forming the constraints (as revealed in Figure 4.13). The results in Figure 4.12 are indicative that for the deterministic procedure, there seems to be no trend in the accuracy of the results for both the variables included and those excluded in the spatial microsimulation process. The results for the stochastic SA method taken alone indicate that accuracies can only be guaranteed for only those variables included as constraints in the SA spatial microsimulation. Within the stochastic procedure (i.e. considering only the SA results), it is observed that the other non-constrained variables are predicted with lower levels of accuracy and precision (than the constrained ones) as seen in Figure 4.13. Comparing the m-IPF and the SA results however, there is a large margin of difference in the results from the two procedures as indicate from Figure 4.12 and Figure 4.13. The TAE results from the m-IPF (50K scale range) are much larger than those for the SA (5K scale range). As subsequently discussed in the next chapter, the TAE does not always form an objective assessor of the quality of spatial microsimulation.

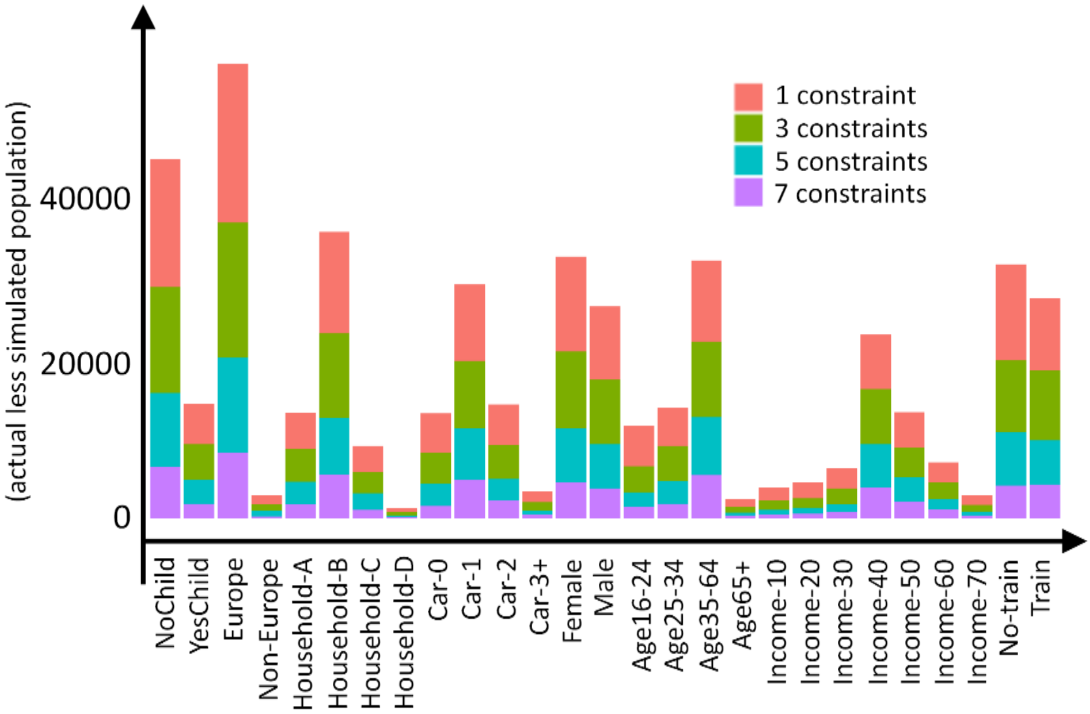


Figure 4.12 TAE vs constraints - deterministic m-IPF.

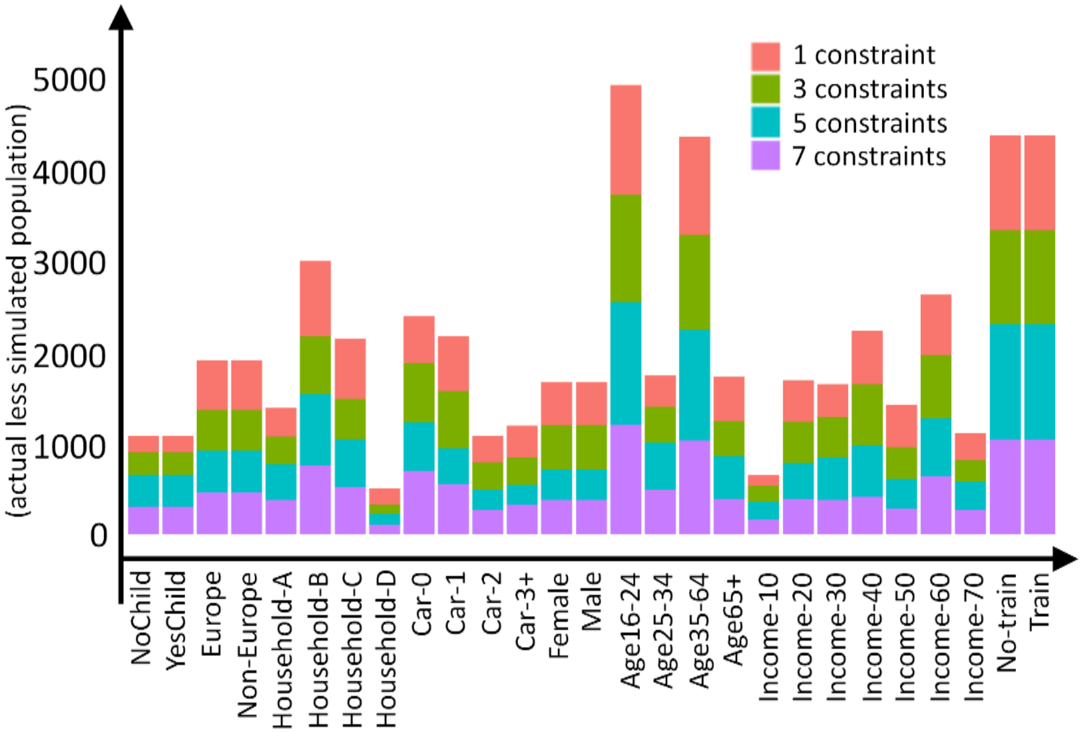


Figure 4.13 TAE vs constraints - stochastic SA.

4.5.2.4 Effect of sample-ratio on accuracy of result

The third experiment investigates the deterministic and stochastic methods to assess the effect of sample ratios of the seed on the accuracy of the TAE results from spatial microsimulation strategies. A Monte Carlo sampling was implemented to choose a range of seed samples of fixed sample ratio. The procedure consisted of selecting samples with a sample ratio v $(0.05\% \leq v \leq 85\%)$ from the full set of NRTS data. This procedure is repeated 250 times for each value of v, and the corresponding 250 TAE values calculated. Density plots shown in Figure 4.14 and Figure 4.15 are created using the 250 estimates for each value of the sample ratio. The Monte Carlo averages out any variability due to the particular choice of sample set. Sample ratios ranging from 0.05% to 85% are used to create the seed, assessing its effect on the TAE for the m-IPF and SA strategies.

The TAE plot of Figure 4.14 shows that the accuracy of the simulated m-IPF population increases with sample ratio, indicated by a leftward drift of the plots as sample ratio increases from 55% to 85%. Further, the increase in density plot heights with sample ratios indicates a nominal increase in precision of the results. A comparison of Figures 4.14 and Figure 4.15 shows that the deterministic m-IPF is more sensitive to changes in sample ratios than the stochastic methods. Like the m-IPF, the SA shows a reduction in TAE values with increase in sample ratios. This is accompanied by a slight reduction in the density levels, implying a reduction in precision (perhaps influenced by an increased variability in larger samples). The stochastic procedures are seen to be less sensitive to changes in sample ratio as indicated by the nominal drift in the plots in Figure 4.15.

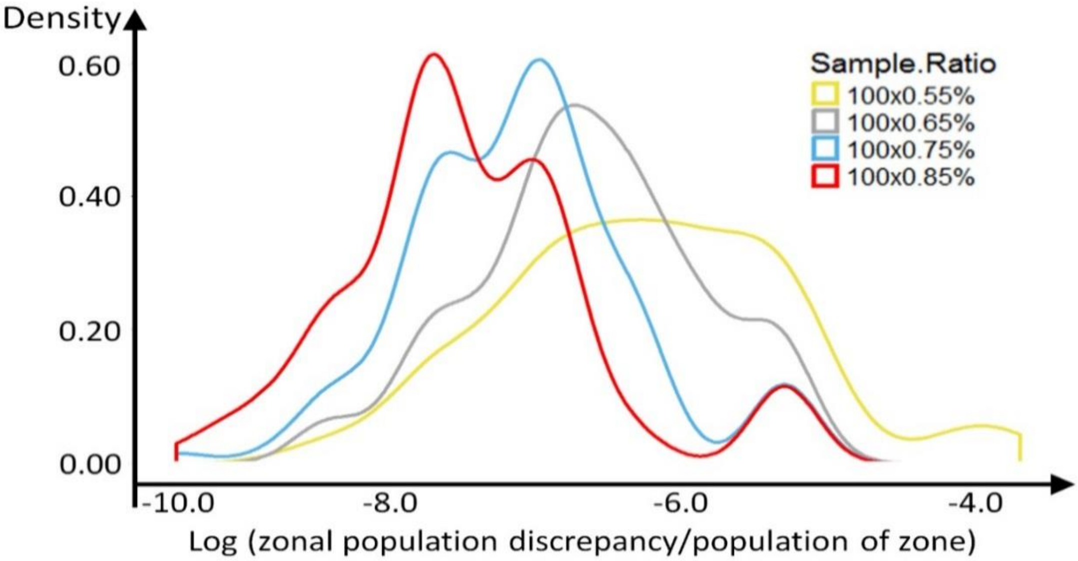


Figure 4.14 TAE vs sample ratio - deterministic m-IPF.

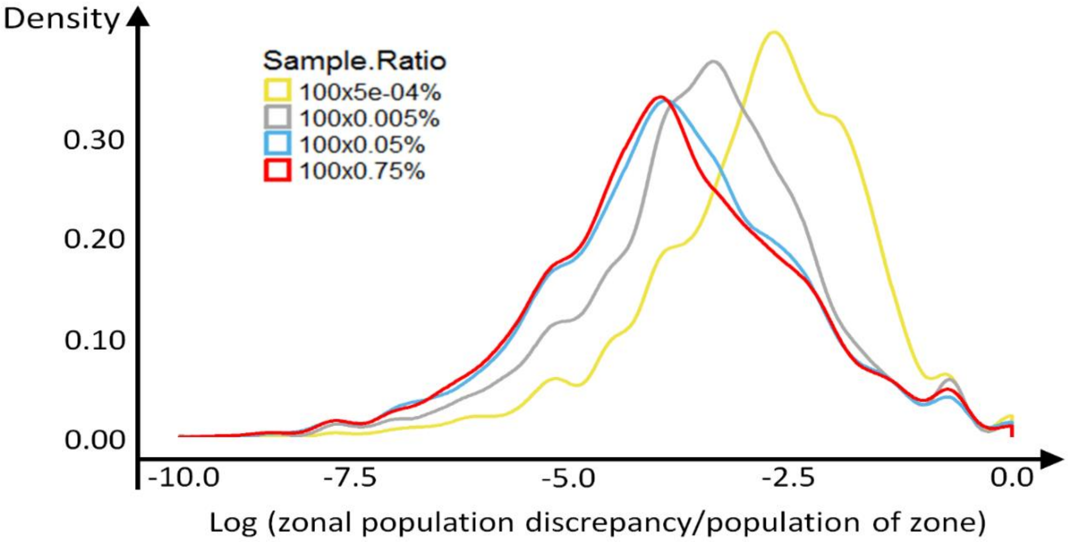


Figure 4.15 TAE vs sample ratio - stochastic SA.

4.5.2.5 Influence of level of skew in seed sample

Traditionally, it is assumed that the disparity between the seed and the aggregate constraint is normally distributed (Ireland and Kullback, 1968, Namazi-Rad et al., 2014). As such, the robustness of deterministic and stochastic strategies need to be assessed for when the seed is skewed, prior to application to typically skewed novel consumer datasets. The NRTS data is stored sorted by origin-purpose and destination-purpose. There are 10 categories[31] for the origin and destination purposes. In this form, a random sample of the NRTS seed variables yields a representative sample. Sampling the NRTS variables with replacement creates a coarsely skewed sample (equivalent to a single bootstrap sample) (Efron and Tibshirani, 1986). Further, selecting the first N tuples of the NRTS would create a strongly skewed seed, as the sample would reveal particular types of passengers reflective of similar travel purposes. A Monte Carlo experiment is conducted, repeating each sampling design about 250 times, and then producing a distribution of the TAE values for the deterministic and stochastic procedures, as shown in Figure 4.16 and Figure 4.17.

[31] The categories of origin and destination purpose are shopping, home, normal workplace, other workplace/meeting, personal business (e.g. doctor, hospital, bank), visiting friends or relatives, sport or entertainment (e.g. concerts, theatre), other leisure activity, school as a student, school accompanying a pupil, taking someone to the airport, station, hotel, meeting someone at a station, airport, hotel, and other purposes not listed.

The sensitivity of the deterministic method to skew in seed is depicted in Figure 4.16, showing the seed progressing from a random sample without replacement, random sample with replacement, dividing the entire NRTS into two halves and then selecting the first $N/2$ tuples (N ranged from 5% to 50% of the entire sample) from each half, and finally, selecting the first N tuples from the NRTS. A similar experiment is conducted using the stochastic SA procedure and the results are shown in Figure 4.17. As seen from the typical results shown in Figure 4.16 and Figure 4.17, seeds that are better representative of the wider population produce in general lower TAE values, indicative of a more accurate simulated representative population. The stochastic procedure however, shows less susceptible to the quality of the seed data. The TAE distribution depicted in Figure 4.17 shows only nominal increases in TAE as the seed becomes less random.

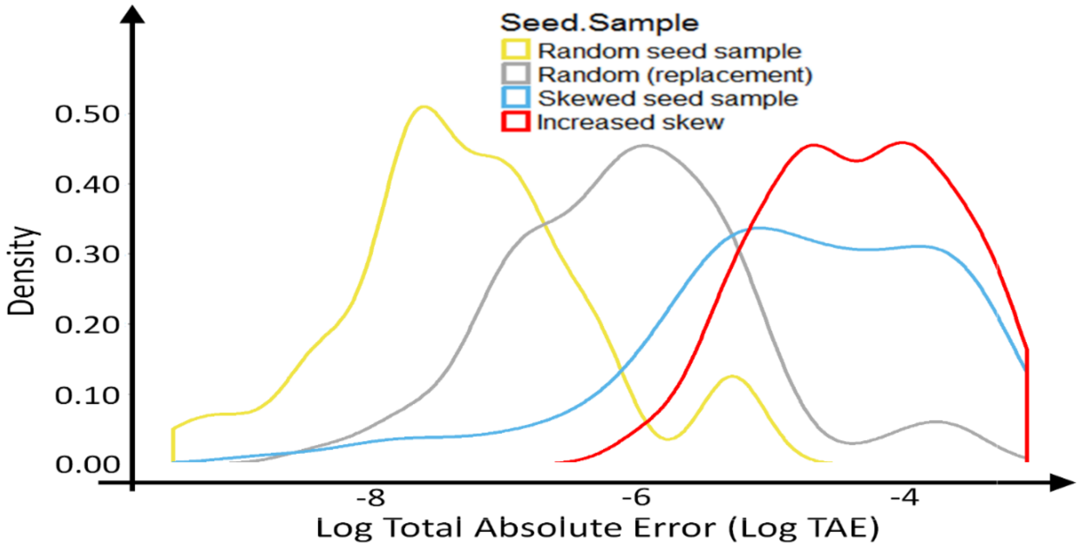


Figure 4.16 TAE vs skew in data - deterministic m-IPF.

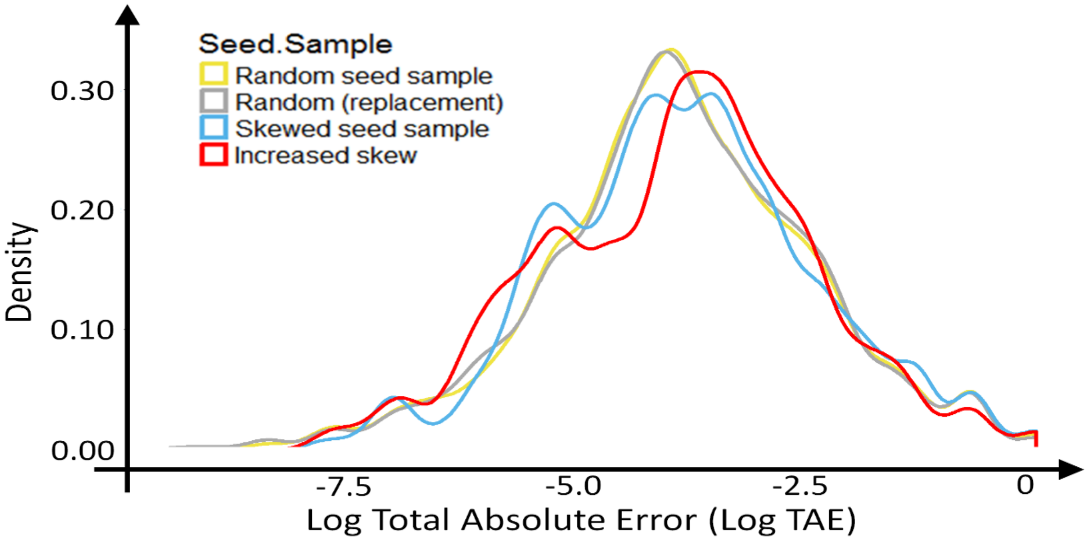


Figure 4.17 TAE vs skew in data - stochastic SA.

4.6 Remarks

The overall intelligence garnered from the experiments indicate that whilst the stochastic procedure yields more accurate results than the deterministic procedure, the SA results are also less sensitive to changes in features of the seed data used in the spatial microsimulation. This lack of sensitivity calls for further assessment of the stochastic procedures to confirm that lower TAE values do actually imply that the distribution of the target is replicated by the simulated population. As shown in chapter 5, lower TAE values do not imply that the distributions of the attribute values of the simulated population are more representative, which would be the desired result in spatial microsimulation. In chapter 5, an m-IPF methodology is developed to enable an exploitation of the full joint distribution of the target population (as opposed to just the marginal used in this chapter), and this forms a more robust basis for comparing the deterministic and stochastic spatial microsimulation procedures.

Re-zonation of the geographic scales from Postcode Sectors to Postcode Areas, or the re-categorizing of variables during spatial microsimulation is sometimes necessary to fit the problem within the computational memory limitations of the R-software (RStudioTeam, 2016). The R-software has a virtual memory limit on the storage size of individual objects as they are stored as characters which are limited to $\mathbf{2^{\wedge}31}$[32] bytes. As such vector, table and array objects cannot exceed this size. As a consequence, a seed table sampling $\mathbf{50,000}$ passengers, from $\mathbf{340}$ zones, with $\mathbf{7}$ income, and $\mathbf{6}$ household type categories would only be able to accommodate another variable of $\mathbf{3}$ categories before exceeding the memory limit for a spatial microsimulation table ($\mathbf{2^{\wedge}31}$). This is the case since $\mathbf{2^{\wedge}31 = 2,147,483,648}$ which is less than $\mathbf{341 * 50000 * 7 * 6 * 3 = 2,148,300,000}$. This illustrates a limitation in using so called 'big data' on the R software platform, and points to areas where computing infrastructure needs to be developed. The R-software code used to implement the deterministic and stochastic experiments are within the CD accompanying this thesis as described in Appendix A.1 to Appendix A.5.

[32] Apart from the object size limitation, the memory available to run a script may also be limited by the users address space limitations or system memory. However, existing high performance computer infrastructure can ameliorate system memory limitations to the tune of about 2 Tera-bytes.

Chapter 5
WEST YORKSHIRE MICROSIMULATION

This chapter takes on from the results of the previous chapter (Chapter 4), where the methodology of spatial microsimulation was presented and the behaviour of deterministic and stochastic spatial microsimulation procedures were investigated. Such investigations enable an assessment of the suitability of each methodology for application to varied datasets which in the case of LENNON ticketing and the NRTS datasets are a skewed subset of the wider population. In this chapter, for the first time, a case study is presented using spatial microsimulation to synthesize an individual-level representative population of railway passengers interacting between geographic zones and through the railway network within the West Yorkshire County. The synthetic data created can be used as input to multi-agent transport models like MATSim (Horni et al., 2016) and SATURN (Hall and Willumsen, 1980), or as presented in Chapter 6 as input to a GIS-GTFS[33] network model. Software like MATSim enable the simulation of individual passenger movement, whilst as shown in this thesis the GIS-GTFS model enables the identification of the context of each passengers mobility, revealing information about actual waiting times endured, the number of stops, crowding in trains, etc. The resulting attribute-rich endogenous and exogenous data from a GIS-GTFS model for instance are useful for subsequent spatial analysis (as presented in Chapter 7 and Chapter 8) to identify particular drivers of mobility interaction on the railway network.

In Chapter 4, the sensitivity and behaviour of the deterministic m-IPF method was compared with those of the stochastic SA method. The influence on the TAE[34] of different types of input seed data was investigated. The m-IPF results were found to be more sensitive to the skew in seed data, when compared with the SA results. In this chapter, a revised multi-dimensional iterative proportional fitting (m-IPF) methodology is developed to address the short-comings of the standard m-IPF adopted in Chapter 4. The revised m-IPF developed in this thesis is shown to be conceptually more robust than

[33] The GIS-GTFS model is a logistical GIS network incorporating detailed transit schedule information (GTFS).

[34] The acronyms are defined in the section on List of Abbreviations.

the newer stochastic simulation-based methods presented in the literature (Farooq and Bierlaire, 2017), and results show the developed m-IPF to outperform the SA method, by producing a simulated population that better reflects the sought distribution of the target population. The revised m-IPF is presented in a practice-oriented way to highlight some of the procedural detail, enabling the results to be reproduced. The dataset combined in the spatial microsimulation implemented in this chapter are rail sector ticketing data from ATOC's LENNON database, the National Rail Travel Survey (NRTS), and the 2011 Census interaction data. The latter two are measured stated surveys, whilst the former are revealed consumer data.

The m-IPF implementation in this chapter consists of a two-stage IPF. The first stage utilizes spatial microsimulation to combine the 2011 Census interaction data with the National Rail Travel Survey (NRTS) producing a first-stage micro-level population. The second stage then combines the micro-population created from the first stage with the LENNON ticketing data. This creates a representative attribute-rich micro-mobility population of railway passengers interacting on the rail network. It is noteworthy to point that the individual-level data are the NRTS and LENNON ticketing data which form biased seed samples because they represent a skewed subset of the wider UK population. The wider UK reference population used are the 2011 Census interaction data.

5.1 Review of the literature

In Chapter 4, the deterministic m-IPF was shown to produce poorer (higher) TAE values, when compared to the stochastic SA method. The stochastic SA method further showed a non-sensitivity to skew in seed data. These results as such would have given the impression that the SA methodology is superior to the m-IPF. It is however, noteworthy to observe that the SA method creates a synthetic population even when there are non-existent seed data for a particular zone. Such an SA simulated population would not be objective. In such a scenario, the m-IPF on the other hand would not simulate any data, resulting in poorer (higher) TAE values for the m-IPF, thereby giving the false perception that SA methods are superior. As such, when non-representative or 'non-existent' data are used, lower $TAE's$ do not necessarily imply a more objective population. The ability of a spatial microsimulation methodology to yield a simulated population with a distribution matching the density (and not just the TAE volume) of the target population is a necessity in simulating an objective population. This in itself raises questions about the efficacy of the TAE as a yardstick for spatial microsimulation convergence[35]. Similar issues concerning the use of the TAE as the yardstick for convergence are reported in the literature (Lovelace et al., 2015, Vidyattama et al., 2011).

In principle, the more objective criteria for assessment of spatial microsimulation would be an ability to yield a simulated population reproducing the full joint (or conditional) distribution of the target population. As such, a better simulated population would be that reproducing the distribution of the target attributes. The methodology presented in this chapter compares and studies the distribution of the simulated and target populations, as a basis for assessing the objectivity of results created from deterministic m-IPF and stochastic SA spatial microsimulation methods. A main criticism of the IPF (and m-IPF) methods in the literature are that they only utilize the marginal information of the target (Farooq and Bierlaire, 2017). As a result, when the sample seed is replicated yielding the simulated population, the heterogeneity in the target population is not reproduced. To address this concern, a methodology is developed in this chapter that exploits the full joint distribution of the target population (as opposed to just the marginal). This is achieved by restructuring the seed data to yield cross-tabulations of variables similar to the cross-tabulation of the target variables. The m-IPF presented below is developed and applied in a practice-oriented way to a case study of railway mobility in the West Yorkshire County.

[35] As reported in the literature, when the seed data are representative of the target population, the m-IPF and SA methods yield comparable results, with differences only dependent on the convergence criteria adopted: total absolute error, percentage classification error, etc. (Duran-Heras et al., 2018, Ryan et al., 2009). However, for consumer data which are typically a skewed sub-set of the wider population (those that consume a particular digital service), the consumer data seed and cannot be assumed to reflect the distribution of the target wider population. Under such conditions, spatial microsimulation developed on the assumption of a representative seed needs to be reviewed (Deming and Stephan, 1940, Kruschke, 2014, Lunn et al., 2012). Some of the TAE results presented in Chapter 4 when the data are skewed, are nonsensical because of the use of TAE as a convergence yardstick. The quality of a spatial microsimulation result should include assessing the density of the simulated population. A simulation more reflective of the distribution of the target population would be a more objective population. In the literature, geographic areas where comparable TAE values were reported for both the m-IPF and SA methods, the seed and the target have had similar distributions (Tanton, 2014).

To facilitate a practice oriented presentation of the revised methodology for m-IPF, details of the process of restructuring the data are presented. This enables a better appreciation of the mechanism for achieving the advantages of data restructuring within spatial microsimulation, as this facilitates an exploitation of the full conditional distribution of the target. It is noteworthy to point out that this section does not serve as a step-by-step manual to the developed procedure; however the concepts are explicitly presented. A guide on the detailed steps could be garnered from the R-script implementation of the developed m-IPF procedure presented in Appendix B, and within the CD disc accompanying the thesis.

Figure 4.3 is a clear visual illustration of the process and mechanism of the m-IPF method. The individual-level seed data in the middle left table includes the individual ID, as well as the 'Age', 'Income' and 'Cars-in-Household' variables. These variables have a combination of 13 variable categories (i.e. 4 for Age, 5 for Income, and 4 for Cars). These are depicted in the re-modelled individual-level seed table depicted in the bottom left of Figure 4.3. If instead of presenting the remodelled table as shown, that the variable categories were cross-tabulated, then instead of having the 13 variable categories, we would have 80 cross-tabulated categories (i.e. 4 x 5 x 4). The re-modelled table would now have categories for instance like $[16-24][Step\_30][Car\_2]$, and there would be 80 potential combinations of these. Referring back to Figure 4.3, the constraint table (shown as the top left table in Figure 4.3) would need to have the same variable categories as the re-modelled seed table to facilitate a reconciliation of the two datasets (i.e. the seed and target). In essence, the target would have 80 distinct cross-tabulated variable categories.

Now referring again to the m-IPF procedure depicted by the rectangular block on the right of Figure 4.3, the available re-modelled seed sample depicted by the bottom left table is shown on one face of the rectangular block as indicated by arrow 3. Similarly, the target data is shown on the top face of the rectangular block as indicated by arrow1. The standard m-IPF procedure would involve iteratively mapping the individuals in the seed to the target volumes, taking each category of variable in sequence, and then for each zone in sequence. As each variable (say Age, Income, Cars, etc.) are mapped in sequence, the standard m-IPF procedure is in effect mapping the seed to the marginal distribution of the target (without reference to the values of the other variables. As such, only the conditional distribution of the target is exploited in the standard m-IPF procedure.

However, by cross-tabulating the variables in the seed, and correspondingly those of the target, the full[36] joint distribution of the target are exploited, yielding a more granular and objective match between the seed and the target. This concept is explored further below in comparing the mechanism of the m-IPF with the stochastic methods. The intuition behind the improved simulation due to cross tabulation of the seed is that the simulated population would consist of individuals who satisfy specific combinations of age, income, and number of household cars. This would be more specific and accurate than an alternative scenario whereby any individual satisfying either the age, or income, or household cars would qualify for inclusion. Further, cross-tabulating the seed has an advantage when dealing with flow interaction data involving an origin and a destination. The cross-tabulation by associating an origin and destination to each individual in the seed ensures the incorporation of the mobility interaction in the microsimulation process.

This methodology also resolves issues which arise when the seed sample replicated to create the demand typically does not have the variety of the target population. This results in the so called ‘‘zero-cell-value problem’ (Guo and Bhat, 2007, Lomax and Norman, 2015). This problem is aggravated by the difficulty in simultaneously controlling for more than one level of constraint (say origin, destination, and individual, i.e. a simultaneous 3-level constraint hierarchy (Müller and Axhausen, 2012)). In this regard, the novelty introduced in this research re-structures the railway ticketing data by converting an ‘individual’ unit in the seed data to an ‘origin-individual-destination’ unit, thereby also better reflecting mobility interaction. This strategy serves as a means of exploiting the conditional marginal distributions (instead of just the marginal) in multi-dimensional target datasets (Farooq et al., 2013). The process of introducing this construct enables the variable categories in the seed to be matched to those in the target, thereby alleviating the zero-cell-values issue. The additional data hierarchies created by restructuring the seed data however, change the dimension of the data table, effectively reducing its length. This alludes to the advantages of using novel big consumer ticketing data which are of considerable volume and variety that a reduction (e.g. 3-fold) in data points would still provide a seed with adequate count and variability.

[36] Strictly, this would be the full conditional distribution (as opposed to full joint distribution) as typically the list of variables available for spatial microsimulation would not form and exhaustive set. Typically not all variables are observed.

The methodology presented in this chapter further serves to explore whether the newer MCMC simulation based methods that employ Gibbs sampling (GS) are revolutionary compared to the established SA and Hill Climbing (HC) stochastic methods (Williamson et al., 1998). These established SA and HC methods are based on the Metropolis-Hastings algorithm. An explanation of the mechanisms of the IPF, GS, SA and HC procedures are presented and formalized, and the conceptual inconsistency in the newer GS methods are reviewed. The methodology developed in this chapter highlights the limitations of the GS methods when the input data are non-representative (i.e. skewed) relative to the wider population. Under such conditions, an imputation (missing data) model would be required for objective simulation using the GS method. Such implementation incorporating an imputation model is currently not available in the literature.

Further particular challenges in creating a representative synthetic demand population have been highlighted in literature (Farooq et al., 2013, Guo and Bhat, 2007, Müller, 2011, Müller and Axhausen, 2012), and some of these have been addressed in recent research. The issues addressed include external validation (Edwards et al., 2011), and non-integers in deterministic strategies (Lovelace and Ballas, 2013). Additional issues raised in the literature about fitting more than one contingency table and scalability with respect to number of attributes (Farooq et al., 2013), have been largely addressed by the multi-dimensional-IPF methods (Barthelemy et al., 2016). Despite the positive reviews of recent MCMC simulation based population synthesis (Farooq et al., 2013), protagonists of established IPF's indicate that it's less-sensitivity to numerical and round off errors, and non-suboptimal solutions make IPF a preferred solution in many complex mobility scenarios.

Following-on from the above review, this chapter sets the foundation for establishing the use of m-IPF methods as the current preferred strategy for spatial microsimulation when the seed sample is skewed relative to the target population. The methodology developed addresses the shortfall highlighted in the literature as detractors of the m-IPF (Farooq and Bierlaire, 2017). The m-IPF is applied to rail sector ticketing data, and the case study of the railways in West Yorkshire presented in this research is the first application of spatial microsimulation to synthesize individual railway passengers from the Census, the NRTS (DfT, 2013a) and LENNON ticketing data. Whilst the Census is representative of the wider population, the NRTS and LENNON are respectively stated and revealed preference data from a skewed subset of the wider population who use the train service.

5.2 Methodology

The study area consists of the railway network in the county of West Yorkshire as illustrated in Figure 5.1. Some aspects of the study area have been earlier introduced (see Section 3.1). The rail sector consumer data for the West Yorkshire study area are harnessed by separately applying the deterministic and the stochastic spatial microsimulation methodologies. This provides a basis for validating both methodologies. As discussed earlier, stochastic or deterministic spatial microsimulation are broadly classed as constrained optimization strategies. The stochastic methods are variants of the MCMC methods, and typically consist of transition chains converging to a solution (Plummer et al., 2006). The m-IPF (Barthelemy and Suesse, 2016, Fienberg, 1970) is currently the predominant deterministic method amongst population geographers, and consists of weighted projections converging to a solution (Ruschendorf, 1995). Protagonists of newer stochastic MCMC simulations are typically transport modellers, and in this section the pitfalls in deterministic and stochastic methods are explained by illustrating their associated mechanisms. This forms the basis for adopting the m-IPF strategy as the choice method, and developing a 3-level hierarchical constraint model to operationalize the process of creating an objective attribute-rich representative micro-level population of railway passengers.

5.2.1 Data and settings

The West Yorkshire inland metropolitan county is made up of five districts in the north of England, with a population of 2.2 million, and covering about 2030km^2. A layout of the railways in the West Yorkshire study area is shown in Figure 5.1. The flows analysed in this research consists of those flows which originate and terminate within the West Yorkshire study area. A large volume (~40%) of railway (National Rail) flows through West Yorkshire either emanate or terminate outside the county (WYCA, 2016), and this affects the volume of passengers that are simulated in this research. In a more realistic setting, the entire volume of tickets captured in the LENNON flows for the entire UK would have been more representative of mobility on the rail network. However, even if that were the case, at this stage of development of computational tools, there would still be the problem of the limitation of the size of data variables and associated tables (2^{31} cells) that could be handled by the analysis software (RStudioTeam, 2016). The inference deduced from these analysis are valid irrespective of the reduced datasets as emphasis is placed more on dealing with the skewed nature of the data rather than on an ability to predict precise passenger demand.

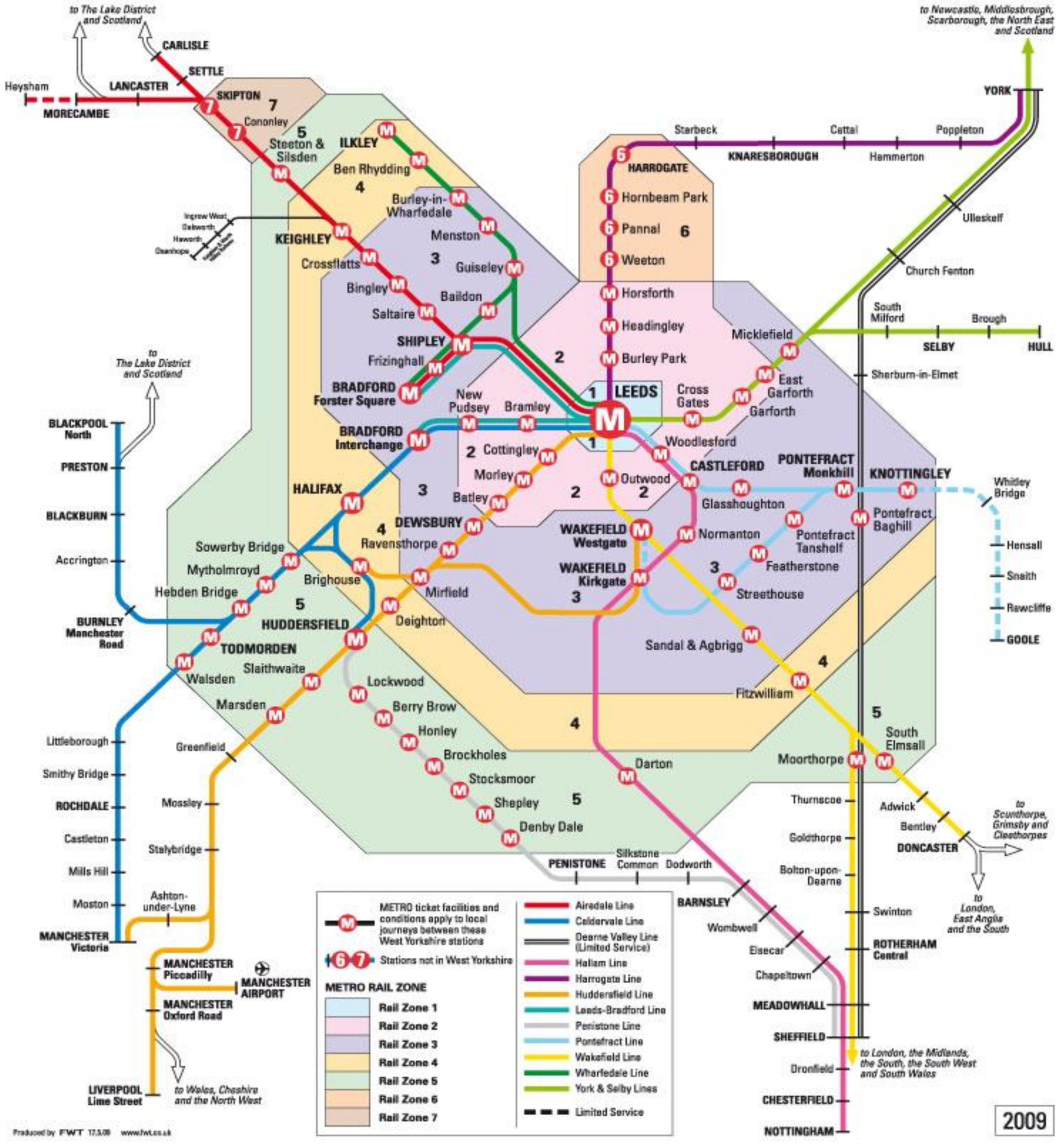
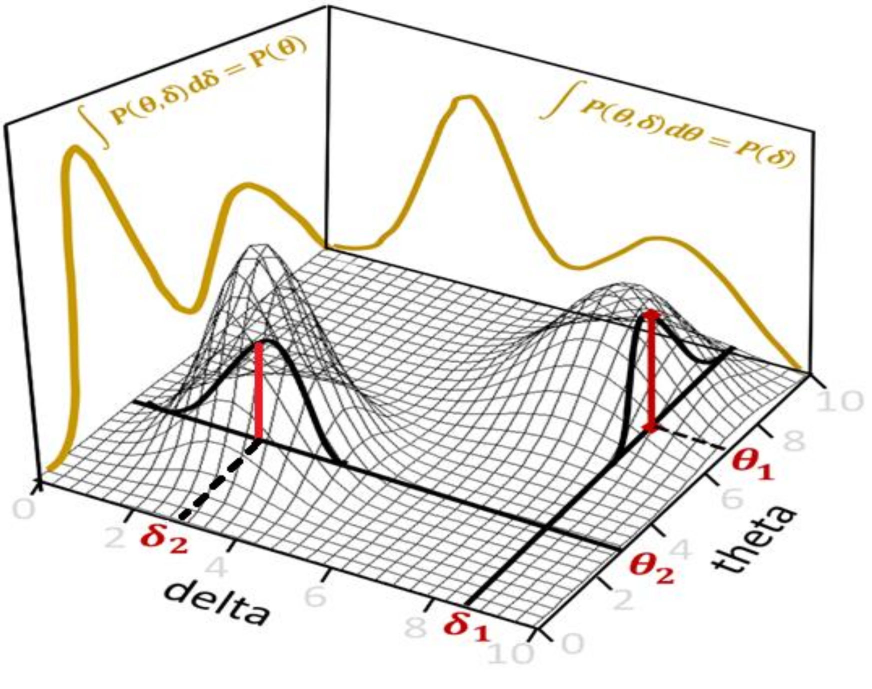


Figure 5.1 Rail network - West Yorkshire (www.wymetro.com)

5.2.2 Mechanism of methods

The mechanism of the stochastic MCMC variant Metropolis-Hastings based (SA and HC), the newer Gibbs sampler based synthesis (GS) (Farooq and Bierlaire, 2017), and the deterministic m-IPF, can intuitively be explained using the Bayesian framework. Details of the Bayesian framework are developed in chapter 7, but here Bayesian concepts are simply used to compare the deterministic and stochastic spatial microsimulation methods. The distinction between traditional statistical (so called frequentists) methods and the Bayesian methods are simple: while creating models representing a phenomena, frequentist methods conceive the system parameters as fixed quantities with confidence intervals defining our certainty in the values, whilst Bayesian methods conceive the system parameters as distributions (density functions) with credible intervals defining the interval of higher confidence in the propensity of solution (Lunn et al., 2012). As such, in the Bayesian framework each system parameter are represented by a range of values with those within the credible interval occurring more often. For example, for a classic spatial interaction model[37] $I_{ij} = O_i D_j e^{\delta d_{ij}}$, in the Bayesian framework, the equivalent distance decay parameter δ (Dennett, 2012) instead of being a fixed point value (say, -1.5 in a frequentist approach), is a distribution. Figure 5.2 illustrates such a parameter distribution surface (also called the posterior). In this instance a two parameter$(\delta,\ \theta)$ model yields a 2-D parameter surface. In the frequentist framework the parameter values would be fixed quantities δ_2 and θ_1 instead of the surface distributions shown in Figure 5.2. The Bayesian notion of parameters being distributions has been found intuitive (Lunn et al., 2012), and the resulting parameter value typically quoted would be the modal (or median) values of the curves on each axis. The credible intervals (CI's) would define regions of higher parameter values which have a higher propensity of representing the phenomena being modelled.



$\int P(\theta,\delta)d\delta = P(\theta)$
$\int P(\theta,\delta)d\theta = P(\delta)$
0
2
4
6
8
10
delta
δ_2
δ_1
0
2
4
6
8
10
theta
θ_2
θ_1

Figure 5.2 Illustration of Gibbs, Metropolis-Hastings and m-IPF.

[37] In the classic expression, I_{ij} represents the interaction or flow between two entities, typically zones, with an origin zone O_i and a destination zone D_j. While δ is a so called distance decay, representing the propensity of individuals to resist travelling longer distances d_{ij} between O_i and D_j.

In the next section the Bayesian concept for model parameter values is used to explain in an accessible manner, the conceptual differences in the deterministic m-IPF, and the range of stochastic (SA, HC, and GS) spatial microsimulation methodologies. Before explaining the deterministic and stochastic methods, the concept of Bayesian analysis is presented to enable subsequent understanding of the Bayesian terminology.

By way of a simple example, imagine a process is represented by points scattered on a plane, and it is desired to derive a unique description or characteristic of the points (perhaps the slope, intercept, and a measure of scatter of points). These unique characteristics reflect the process which generated the points. The aim then would be to deduce an optimal line that best describes the range of points. In the Bayesian framework, the slope and intercept[38] are conceived as distributions (typically like that shown in Figure 5.2) to reflect that the process which generated the data has stochastic variation. The mode of these distributions describe parameter values of higher propensity, as shown in Figure 5.2 whereby θ might represent the slope and δ the intercept with modes at the peaks near θ_1 and δ_2 respectively (as shown in Figure 5.2).

In the Bayesian methodology, a model is created for the data by specifying a likelihood function. For example, if we can assume that the outcome data points are randomly scattered, then the likelihood function could appropriately be a normal distribution. In the alternative for example, if the outcome variable were count data, then perhaps a Poisson or negative Binomial likelihood function would be appropriate. The model created for the data is called the 'likelihood' (or more specifically the likelihood function). Having created a model for the data, we secondly create models for each of the parameters (in this case the slope and intercept). As the points are conceived to be randomly scattered around a line, a model for the slope could be a normal distribution, and perhaps as the intercept tends to be a fixed constant, we can conceivably model it as a flat or uniform distribution. Other models could be adopted for the parameters based on our intuition and knowledge about the process which generated the data (Lunn et al., 2012). The models for each of the parameters are called 'priors', as they represent our prior belief about the parameters.

[38] Note that the Bayesian framework adopts the MCMC algorithm to solve a wide range of optimisation problems. The slope and intercept are parameters of the conceived model requiring an optimum solution.

In the Bayesian framework, based on the classic Bayes rule (Bayes et al., 1763), the likelihood and prior models are multiplied to give the qualitative distribution of the posterior. In practical applications, the MCMC procedure is adopted to deduce an optimal distribution of the posterior yielding the parameter distribution shown in Figure 5.2. The MCMC optimisation procedure consists of first specifying an initial proposal distribution, and then an initial parameter value is sampled from this distribution. A new posterior distribution is deduced (by multiplying the prior parameter model by the data likelihood function at the initial parameter value). Secondly, a new sample is taken from this new posterior parameter distribution. The initial proposal and the new posterior parameters are compared and a choice of which one to keep is decided such that higher parameter values have a higher probability of being retained. The criteria for making this choice distinguished the different MCMC[39] procedures, but in any event the procedure described above is repeated recursively. As higher parameter values have a higher propensity of being chosen, the parameter surface created ensures that the distribution peaks at the modal parameter value.

The intuition behind using the MCMC methods, and in particular in the Bayesian framework is simple: the essence of statistics analysis is to summarize the phenomena represented by a dataset, and to make inference based on these (Fisher, 1922). In complex scenarios, a closed form expression cannot always be derived for the models that represents the data. Under such conditions the MCMC algorithm can be used in estimating optimal parameter values which represent optimal solutions for such complex models. Once a model is created by the practitioner or data analyst, to represent a particular objective function being optimized, the MCMC algorithm can be used to deduce parameter values for such a function.

The Monte Carlo Markov chain (MCMC) algorithm is in essence a two prong process: an exploratory and an optimisation process. The repeated sampling process (the Monte Carlo) enables an exploration of the solution space, and the criteria and steps in deciding the choice of posterior values to retain in order to proffer the best solution is the optimisation process.

[39] More technical descriptions of the MCMC procedure are widely available in the literature, but here we present an accessible description simply for brevity.

5.2.2.1 IPF in the Bayesian framework

We shall now use Figure 5.2 to illustrate and succinctly explain mechanism of the classic IPF algorithm: In the IPF, the posterior parameter space is projected onto one of the parameter planes as shown by the brown curves in Figure 5.2. The δ plane forming the marginal distributions $P(\delta)$ is created by integrating over θ, i.e. $\int P(\theta, \delta)d\theta$. In implementation, the mechanism of the IPF procedure is such that the distribution of the sample seed $(x \in X)$ is multiplied by a weight distribution to project it onto the marginal distribution, creating a fit and a new proposal distribution. This new proposal is then weighted and then projected onto the second plane, the θ plane, creating a new fit on the marginal parameter $P(\theta) = P(\theta, \delta)d\delta$. Each projection is achieved by creating the appropriate weight to enable the seed to iteratively fit the $P(\delta)$ and $P(\theta)$ marginal distributions. The process is iteratively repeated until convergence whereby the weight matrix reaches an equilibrium or does not change further beyond a pre-defined convergence criteria. The particular criticism of the IPF is that it aims to fit the seed data to marginal distribution of the target population (i.e. the brown curves in Figure 5.2), and not to the full joint distribution of the target (represented by the full mesh). As a result the simulated population lack in the full variability of the target. The m-IPF strategy developed in this chapter aims to resolve this problem by restructuring the seed data to match the cross tabulation of the target data, thereby reflecting a fit to the full joint distribution of the target data.

The IPF method is dependent on multiplying a proposal distribution by a weight to project it onto a marginal (or as is the case in this research onto a full or conditional) posterior distribution. The proposal distribution is constituted from a sample (seed) from the population. As such, in areas or zones where a particular seed does not exist, no weight can be created, and no simulated population is created for that zone. In essence, for the IPF to yield accurate results, the proposal distribution would need to be reflective of the target population. This aspect of the IPF can in fact be seen as an advantage as it ensures that no simulation results are created, instead of creating results that are absurd[40].

[40] When the seed is non-representative of the target population, the stochastic HC and SA methods still yield a simulated population. Such results distinguish between population reconstruction and population synthesis. In reconstruction, the simulated population is created arbitrarily in the proviso that the marginal values of the target are fulfilled.

5.2.2.2 HC/SA in the Bayesian framework

There are two established MCMC samplers, the Metropolis-Hastings and the Gibbs sampler. The Metropolis algorithm is akin to the spatial microsimulation methods of Hill Climbing (HC) (Williamson et al., 1998) and Simulated Annealing (SA) (Kavroudakis, 2015) used for population synthesis. The Metropolis algorithm can be described as a random walk in the parameter space (Kruschke, 2014). When applying the HC and SA for synthesis, a proposal sample of size equal to the population being simulated is taken from the population to form the seed. Parameters of this proposed seed, say the marginal aggregate volume of people aged 65+, could represent a point (along the θ-axis shown in Figure 5.2) in the parameter space. If this sample point (the aggregate volume) is equal to the marginal volume of people aged 65+ in the actual population of the zone, the point is accepted. Realistically however, the population would have many more parameters other than just say the aged 65+. The combination of these parameter values would create the parameter space $\mathbb{R}^N$. Where the 'N' represents the dimension of the parameter space. In practice the Metropolis-Hastings algorithm recursively samples and updates the initial proposal distribution, forming the true posterior parameters space which satisfies the target parameter constraints. The difference in the variants of the Metropolis algorithm (Simulated Annealing, Hill Climbing, etc.) are in the criteria for deciding when to reject the parameter values that don't exactly match or optimize an objective function yielding the target.

The Metropolis based HC and SA methods suffer similar issues to IPF methods in that the proposal distribution has to be reflective of the target distribution. In practice for the HC and SA methods to be efficient, a strategy needs to then be additionally developed for tuning the proposal distribution to that of the target. As widely reported in the literature (Kruschke, 2014, Lunn et al., 2012), if the proposal distribution is too narrow or too broad, an immeasurable number of steps (chains) may be required to adequately explore the full parameter space, with the attendant possibility of the trajectory being stuck at a local optimum (Kruschke, 2014, Williamson et al., 1998). Due to the general and broad nature of the Metropolis based synthesis methods, any proposal distribution would yield result, albeit sub-

In synthesis on the other hand, a zonal population created emanates from replicating samples associated with that particular zone, excluding any sample associated with other zones.

optimal and biased. This may at times create the false impression that the HC and SA are better methods for yielding a result albeit absurd, when methods like the IPF would as an inherent precautionary measure not yield any results. These issues were intimated earlier (see Section 5.1), and bring into question the sole use of TAE as a yardstick for assessing the quality of spatial microsimulation results. A more robust yardstick would bring into consideration the ability of the population synthesis method to replicate the full-joint or conditional distribution of the target population. These issues are highlighted in this chapter in validating the various deterministic and stochastic spatial microsimulation methods.

5.2.2.3 GS in the Bayesian framework

The Gibbs sampler (Geman and Geman, 1987) is designed to be a random walk on the parameter space like the Metropolis (Hastings). The difference in the Gibbs sampler (GS) is that the parameters are dealt with one at a time (in series) when exploring the parameter space, as opposed to the Metropolis algorithm where random samples of the seed (reflecting the full joint parameter distribution) are considered jointly in-parallel. In implementing the GS method (referring again to the 2-D parameter space of Figure 5.2), arbitrary values δ_1 and θ_1 are initially chosen for the parameters. The value of θ_1 is first updated by sampling from the posterior distribution conditional on δ_1. The conditional distribution is the parameter curve created by slicing through the joint parameter posterior at point δ_1 (i.e. $p(\delta_1|\theta_1, X)$ shown in Figure 5.2). Randomly sampling the resulting conditional parameter curve yields an update of θ_1, a point on the curve θ_2. Secondly, the value of δ_1 is updated by sampling from posterior distribution conditional on θ_2, thereby yielding the point δ_2 (i.e. $p(\delta_1| \theta_2, X)$). This process is repeated each time updating the θ_i and δ_i for $(1 \leq i \leq \infty)$, thereby gradually simulating the posterior parameter surface. In practice, as the joint posterior is unknown, the conditional distribution is derived from a model of the covariates, and the parameters of the population are deduced by sampling from the posterior predictive distribution.

The GS has similarities to the IPF in that the procedure does not suffer the inefficiency of rejected proposals as is the case with the Metropolis based algorithms. As such for well-mannered problems (Roberts and Rosenthal, 2004), the IPF and GS tend to converge to a solution quicker and hence more efficiently. The particular challenge in adopting the GS method is that the analyst must have an ability to deduce the conditional probability for each parameter, on all other parameters to enable posterior samples to be generated. In scenarios where a set of the data used is a skewed representation of a wider dataset, for example when the NRTS and LENNON data are used in conjunction with the Census, then a missing data model needs to be created to explain the difference between the skewed and holistic datasets. Such a missing data model cannot always be readily created without knowledge of the statistical mechanism through which the data was created and the missing values became non-available in the first instance. This issue is elaborated on further in the next section as a particular impediment to recent GS simulation based population synthesis methods (Farooq et al., 2013). Protagonists of such simulation based GS methods (Farooq and Bierlaire, 2017, Grapperon et al., 2016) have heralded such methods as being superior to m-IPF methods, even in scenarios where the datasets combined are skewed and non-representative of the wider datasets. The next section is important as it highlights how such an assertion of the superiority of GS methods in skewed data scenarios is conceptually implausible. The section presents a formalisation of the claim to the contrary.

5.2.3 GS population synthesis

In the case of consumer data like ticketing data, which are typically skewed relative to the wider population, the skew implies that the data cannot be assumed to stem from a randomized control survey of the wider population. Inclusion of the skewed data with data from a wider population in an MCMC design matrix yields a so called missing-not-at-random (MNAR) scenario, whereby the disparity between the skewed data and the wider population is systematic. In such a case, a model has to be incorporated to reconcile the skewed distribution with that of the wider population. Otherwise inferences drawn from an analysis which directly combines the two disparate distributions would not be conceptually consistent, as the two datasets represent different phenomena and attendant data generation processes. To demonstrate this formally, let the sample seed from the wider population be represented as X, which in turn can be conceived as partitioned into an observed part X_{OBS} and the missing part X_{MIS}, whereby $X = (X_{OBS}, X_{MIS})$. Note the observed part corresponds to the subset typical of skewed consumer data. If M is defined as the matrix of ***missingness***[41] with same dimension as X with elements 1 or 0 depending on whether a corresponding

[41] 'Missingness' was originally a colloquial term that relates to the statistical mechanism that resulted in data values being missing, and refers to the pattern of distribution of the variables with missing data values.

element in M was observed (1) or missing (0). When data are missing, the probability model to describe data is the full joint probability given by $P(X_{OBS}, X_{MIS}, M|\theta,\varphi)$. Where θ represents the parameters of the phenomena being modelled, and φ the parameter of ***missingness*** (Rubin, 1978), which describes the statistical missing data mechanism.

Integrating over the missing values X_{MIS}, gives:

$$\int P(X_{OBS}, X_{MIS}, M|\theta,\varphi)\ dX_{MIS} \\ = \int P(M|X_{OBS}, X_{MIS},\varphi)\, P(X_{OBS}, X_{MIS}|\theta) dX_{MIS} \tag{6}$$

Observe the first term on the right (after the equal sign and under the integral): It is the distribution of M, that is $P(M|X,\varphi)$. Assuming that the data are statistically missing at random (MAR) (Rubin, 1978), then the first expression within the integral on the left of Eqn. (6) simplifies to $P(M|X,\varphi) = P(M|(X_{OBS}, X_{MIS}),\varphi) = P(M|X_{OBS},\varphi)$, as by definition the '***missingness'*** depends only on the observed data (Little and Rubin, 2002, Schafer, 1997). If in addition to the data being statistically missing at random (MAR), that further φ and θ are distinct, such that the mechanism that yields the missing data is distinct from the mechanism being inferred: A practical example would be a scenario whereby the season ticket journey factors (rate of ticket use) are not known simply because there are not enough sensors on the rail network to capture and register each passengers repeat journey. These missing journey factors are distinct from the drivers of mobility phenomena being inferred in the analysis of ticketing data. As such, it can be said that the missing data mechanism for season ticket journey factors φ are distinct from the mobility mechanism being inferred θ.

With φ distinct and as such independent of the LENNON data parameter θ, the full joint distribution reduces to:

$$P(X_{OBS}, M|\theta,\varphi) = P(M|X_{obs},\varphi) \int P(X_{OBS}, X_{MIS}|\theta)\, dX_{MIS} \tag{7}$$

$$P(X_{OBS}, M|\theta,\varphi) = P(M|\varphi, X_{OBS})\, P(X_{OBS}|\theta) \tag{8}$$

$$\therefore \quad P(X_{OBS}|\theta) \propto \mathcal{L}(\theta|X_{OBS}) \tag{9}$$

The above is the case, as the '***missingness'*** is only dependent on the observed data, and $P(M|\varphi, X_{OBS})$ can be seen as a normalising constant.

Based on the above, the likelihood of θ, the parameter of mobility, given only the observed data is equivalent to the full joint probability, and under such circumstances inference based on θ would be valid. This deduction assumes that the data are MAR, and that the missing data mechanism is independent of the mobility mechanism being inferred. However, as the consumer ticketing data are skewed (biased), it is not MAR and valid inference regarding the wider synthetic population (i.e. the full joint distribution) can as such not be deduced from a combination of the skewed seed consumer data with data from the wider population within the Gibbs MCMC model framework. Skewed consumer data are an MNAR subset, and requires an imputation model to be built prior to or integral with any Gibbs based MCMC simulation based population synthesis strategies. The imputation model would account for the difference between the skewed consumer data attribute values and those of the wider population.

The recent proposed MCMC simulation based population synthesis (Farooq et al., 2013) are based on using partial views of the available joint distributions of agent's attributes to create a mobility model. This model forms the basis for simulating draws from the target distribution. However, in the case where the agent population is skewed, the resulting covariates (pertaining to only these agents) are skewed and not a random subset of covariates from the wider population. As such there are no conceptually consistent bases for combining agent attributes with marginal or conditional attributes from the wider population. As a result, claims of the superiority of simulation based spatial microsimulation over IPF strategies as reported in the literature (Farooq et al., 2013, Grapperon, 2016) are unfounded in the current form. What is required is to incorporate an imputation model within the more traditional model used to derive the conditional probability of each parameter on the others. This would then form a basis for generating samples from the conditional distributions. Such an implementation can be fulfilled within the Bayesian framework (Lunn et al., 2012) which enables the simultaneous incorporation of an imputation model within an analysis model.

To the best of our knowledge, such a simulation based population synthesis that incorporated a missing data model has not been reported in the literature, and for expedience is beyond the scope of this research. As such, based on this conceptual shortfall, the MCMC simulation based population synthesis is not investigated further, but is further discussed at the end of the thesis as anticipated and scheduled for future work.

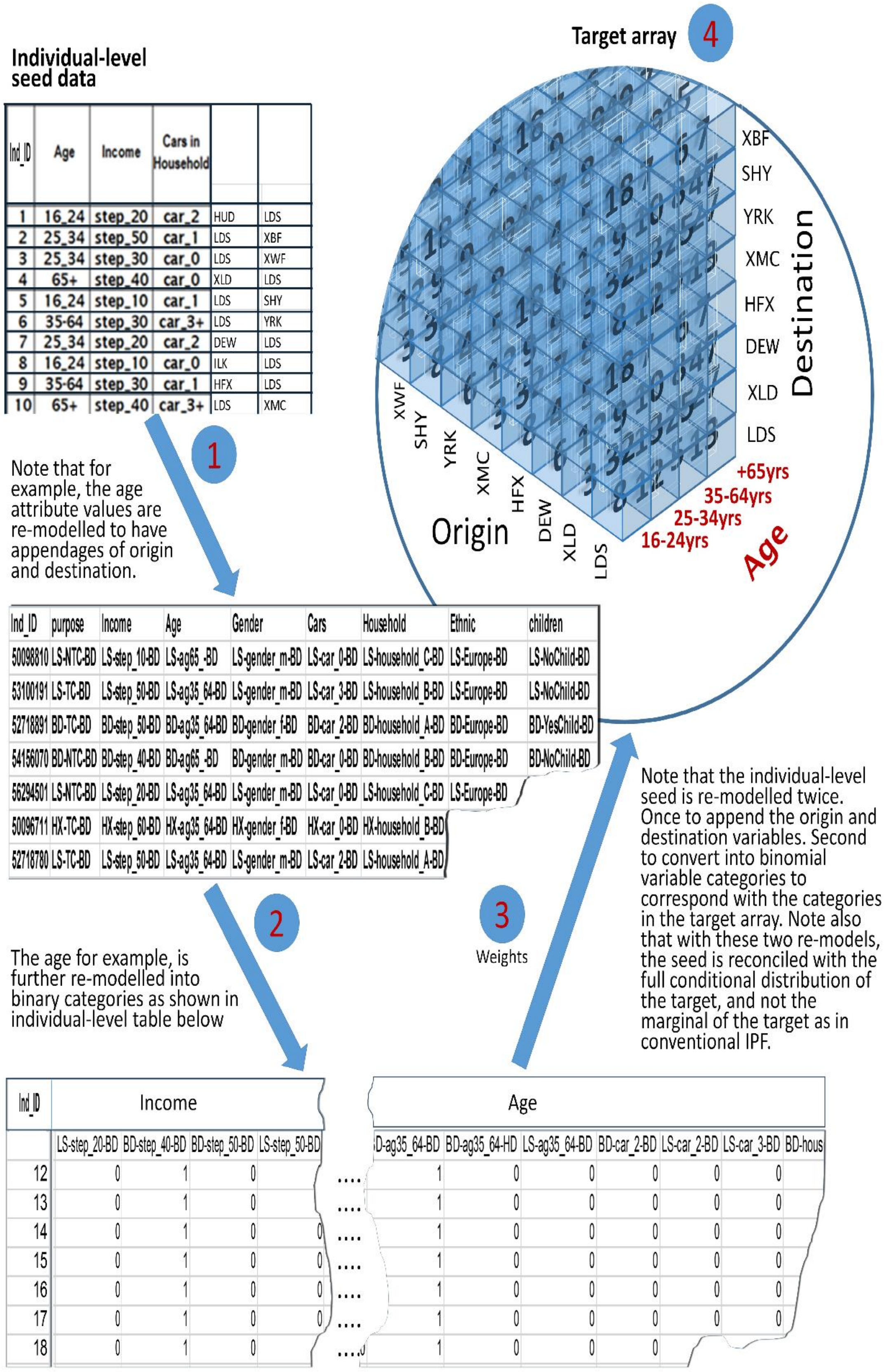
5.2.4 Procedures for implementation

The first hint in the literature about the potential of microsimulation for use in interaction data dates back to 1968 (Ireland and Kullback, 1968), but to date no application has been reported in literature which include the origin-destination (O-D) interaction as an explicit variable constrained simultaneously along with the other variables of mobility in the procedure. To ensure that the identity (ID) of the passengers are carried along, the ID information is included in the seed sample, but during spatial microsimulation, this information is not subject to constraint, thereby producing a spatial micro-simulated[42] population with the distribution of the ID's included. It is noteworthy to point out that the attachment of each passenger's particular origin and destination to the passengers attribute and the categories therein effectively incorporates the mobility interaction in the spatial microsimulation process. To date this is the first implementation of such a hierarchical constraint model, and this marks this work out as a pioneering spatial micro-mobility implementation. The methodology developed ensures that the full distribution within the target cross-tables are exploited (as opposed to only exploiting the marginal).

Each individual (for example of age25-34) is associated with a unique origin (out of the N many origins) and a unique destination (out of M destinations), additional variability is inherently introduced to that ag25-34 category by appending an origin and destination pair. In essence the single attribute category ag25-34 now has N by M potential variations to it, and so also does all the other attributes and their categories. The attachment of origin and destination pairs to each variable category value in the seed yields a 3-level simultaneous hierarchical constraint. As a result, the spatial microsimulation optimizes a set of 3 attributes simultaneously. This is illustrated in Table 5.1 showing the re-modelling of the individual-level seed, and the match to the target array. In effect, the objective function is equivalent to the cross-tabulated target distribution, instead of the marginal in standard IPF. This yields a synthetic population better following the target distribution.

[42] We make the distinction between spatial microsimulation and population re-construction, whereby the former includes an identification of each individual in the micro-data and the latter creates a constituent volume of the population with no individual identities. For the purposes of mobility on the railways, it is essential that the more demanding S-m-s is implemented and this chapter highlights the particular methodology for achievement.

Table 5.1 Table illustrating spatial microsimulation with the incorporation of multi-level simultaneous seed hierarchies.



Individual-level seed data

Ind_ID	Age	Income	Cars in Household		
1	16_24	step_20	car_2	HUD	LDS
2	25_34	step_50	car_1	LDS	XBF
3	25_34	step_30	car_0	LDS	XWF
4	65+	step_40	car_0	XLD	LDS
5	16_24	step_10	car_1	LDS	SHY
6	35-64	step_30	car_3+	LDS	YRK
7	25_34	step_20	car_2	DEW	LDS
8	16_24	step_10	car_0	ILK	LDS
9	35-64	step_30	car_1	HFX	LDS
10	65+	step_40	car_3+	LDS	XMC

Note that for example, the age attribute values are re-modelled to have appendages of origin and destination.

Ind_ID	purpose	Income	Age	Gender	Cars	Household	Ethnic	children
50098810	LS-NTC-BD	LS-step_10-BD	LS-ag65_-BD	LS-gender_m-BD	LS-car_0-BD	LS-household_C-BD	LS-Europe-BD	LS-NoChild-BD
53100191	LS-TC-BD	LS-step_50-BD	LS-ag35_64-BD	LS-gender_m-BD	LS-car_3-BD	LS-household_B-BD	LS-Europe-BD	LS-NoChild-BD
52718891	BD-TC-BD	BD-step_50-BD	BD-ag35_64-BD	BD-gender_f-BD	BD-car_2-BD	BD-household_A-BD	BD-Europe-BD	BD-YesChild-BD
54156070	BD-NTC-BD	BD-step_40-BD	BD-ag65_-BD	BD-gender_m-BD	BD-car_0-BD	BD-household_B-BD	BD-Europe-BD	BD-NoChild-BD
56294501	LS-NTC-BD	LS-step_20-BD	LS-ag35_64-BD	LS-gender_m-BD	LS-car_0-BD	LS-household_C-BD	LS-Europe-BD	
50096711	HX-TC-BD	HX-step_60-BD	HX-ag35_64-BD	HX-gender_f-BD	HX-car_0-BD	HX-household_B-BD		
52718780	LS-TC-BD	LS-step_50-BD	LS-ag35_64-BD	LS-gender_m-BD	LS-car_2-BD	LS-household_A-BD		

The age for example, is further re-modelled into binary categories as shown in individual-level table below

Note that the individual-level seed is re-modelled twice. Once to append the origin and destination variables. Second to convert into binomial variable categories to correspond with the categories in the target array. Note also that with these two re-models, the seed is reconciled with the full conditional distribution of the target, and not the marginal of the target as in conventional IPF.

Ind_ID	Income					Age						
	LS-step_20-BD	BD-step_40-BD	BD-step_50-BD	LS-step_50-BD		D-ag35_64-BD	BD-ag35_64-HD	LS-ag35_64-BD	BD-car_2-BD	LS-car_2-BD	LS-car_3-BD	BD-hous
12	0	1	0		....	1	0	0	0	0	0	
13	0	1	0		....	1	0	0	0	0	0	
14	0	1	0	0	....	1	0	0	0	0	0	
15	0	1	0	0	....	1	0	0	0	0	0	
16	0	1	0	0	....	1	0	0	0	0	0	
17	0	1	0	0	....	1	0	0	0	0	0	
18	0	1	0		....	1	0	0	0			

A number of practical issues arise as a result of introducing multi-level hierarchies, and these are summarised below so that necessary precautions can be taken in practical applications.

- Numerical approximation errors are introduced during the process of re-zoning (re-scaling) from one geographic boundary definition to another (say from Postcode to MSOA zones). A practical remedy would be to convert the marginal values to percentages, forcing each marginal to sum to one. Subsequent multiplication by zonal populations would yield the individual weights.
- An increase in granularity by creating a 3-level hierarchy in the seed requires a commensurate increase in volume of the seed (creating a compelling case for big data). In essence, with a more complex finer-grain seed, there is an increased uncertainty in the ability to match the full variability of the population, thereby requiring a larger seed volume to sustain the accuracies. An N by M hierarchy effectively reduces the seed by N by M, commensurately affecting the accuracy and precision of the results due to effectively reduced sample ratios.
- The multi-level hierarchy increases the propensity of a mismatch in number of variable categories between the seed and target data. In the West Yorkshire case, there were ~2192 Census variable categories, and ~1351 seed categories. In such a scenario the m-IPF synthesizes lower population volumes and the stochastic procedures instead create full volumes, albeit sub-optimal.
- Counts (of mobility in this case) inherently create sparse contingency matrices. A further hierarchy of constraints exasperates this sparsity. A practical remedy to the zero inflated contingency matrices in spatial microsimulation is to add a small value typically of the order of the error tolerance of the problem (Lomax and Norman, 2016).

In implementing the precautions listed above, each particular scenario is dealt with separately. For example, adding small values to sparse contingency matrices may introduce mass to structural zeros, equivalent to adding individuals to the mobility, effectively altering the dynamics. Matching the categories of the seed and target identifies structural zeros. Addition of small amounts where sampling zeros occur introduce cross-relations between attributes of individuals, and this is managed using a sensitivity analysis for monitoring and to be kept such impacts to on the mobility dynamics to a minimum. Such impacts on the dynamics should be kept to the order of magnitude of the convergence tolerance.

5.3 Results

In developing the methodology (Section 5.2.3), a formalisation of the statistics theory was presented showing that the Gibbs MCMC (GS) simulation based population synthesis (Farooq et al., 2013) requires an imputation model to enable application to situations where the seed is skewed relative to the target. As a result the GS methodology was not developed further in this thesis. In Chapter 4, the Hill Climbing (HC) and Simulated Annealing (SA) spatial microsimulation strategies were shown to yield better (lower) TAE values when compared with the standard IPF strategy which relates seed data to the marginal distribution of the target data. In this chapter, the IPF methodology is developed to exploit the full-joint distribution of the target (instead of just the marginal), and applied to simulate a population of railways passengers in the West Yorkshire study area. The datasets combined are the 2011 Census interaction data, the NRTS survey, and the LENNON ticketing data.

Results are presented for the first spatial microsimulation combining the Census information data with the NRTS, revealing the results produced using the improved m-IPF strategy, and corresponding results from the stochastic SA method. The results are then presented for the second spatial microsimulation which combines the simulated population from the first spatial microsimulation stage (i.e. the simulated population resulting from combining the NRTS with the Census), with the LENNON ticketing data. This second spatial microsimulation creates an attribute rich micro-level population of railway passengers embedded in the wider population. As opposed to quoting TAE values, the density of the simulated populations are presented, enabling comparison with the density of the seed, and that of the target population. The appropriate choice of density plot for comparison of the distributions from one synthetic population to another was the normalised density plot with a total unit (1.0) area. The horizontal axis was chosen to be indicative of the number of variables involved in the spatial microsimulation, with a density curvature to holistically distinguish the influence of the individual variables within the synthetic population. The density plot typical of that in Figure 5.3 was chosen, with a detailed description of facets of the plot included in the results section. Following the density results, as a form of validation, query results of the synthetic micro-level cross-tabulated data are presented, postulating that the intuitive nature of the results is indicative of the success of the spatial microsimulation process. These results are further discussed in the context of the wider project.

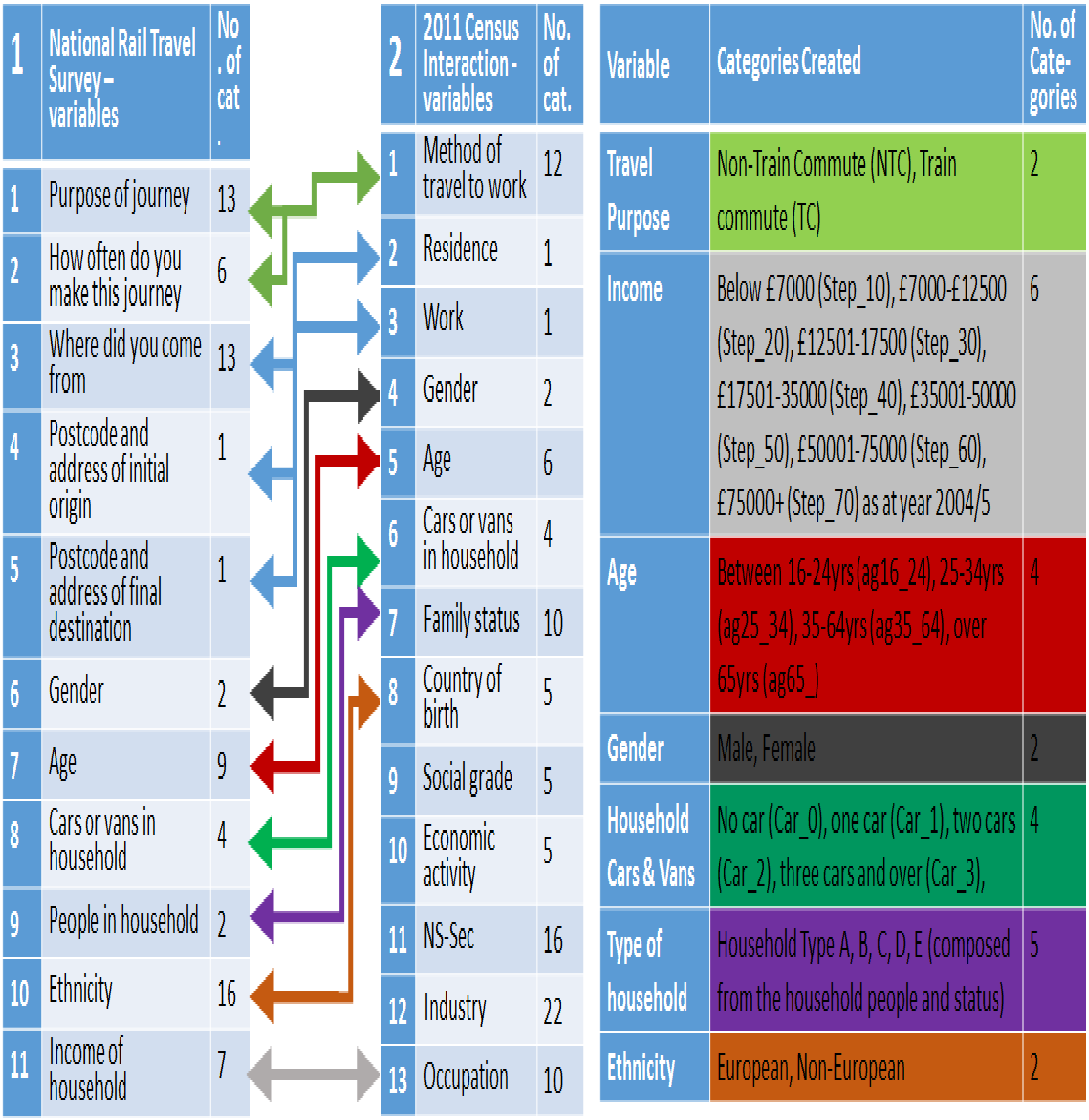
5.3.1 Stage 1 - Linking the Census and NRTS

In the re-developed m-IPF, the cross tabulation in the target data are exploited by re-structuring the seed by associating an origin and destination to each variable-category. Results presented in Figure 5.3 are density functions of the simulated population proportions within each variable category. As there are 8 variables (travel purpose, household type, income etc.) the span of the horizontal axis would be 0.125 (i.e. 1/8), reflective of the granularity in the scale or level of detail resulting from the large number or set of data variables. Note however that the total area under the density plot would always sum to unity to enable comparison of results from different spatial microsimulation exercises. If for instance there were 20 variables used in the spatial microsimulation, then the span of the horizontal axis would be 0.05 (i.e. 1/20). In the results presented in Figure 5.3, there are two categories (NTC and TC) in the travel purpose variable. In the Census, the majority of the population consists of non-train commuters (NTC), with a smaller proportion of train commuters (TC). The high disparity in the simulated population proportions of NTC and TC yields that the 'travel purpose' variable would contribute a substantial amount at extreme (near 0.125) of the horizontal axis of the density plot. As seen in Figure 5.3, the density plot of the Census has a higher value at toward the 0.125 end of the horizontal axis, reflective of the large disparity in NTC and TC proportions. The population created by the deterministic m-IPF is seen have a similar density to the Census. This is reflective of the fact that the deterministic m-IPF yields a simulated population made up of a larger proportion of NTC passengers. Since the Census formed the target in this simulation, the inference then is that the deterministic m-IPF procedure with a hierarchy of cross-tabulated seed data yields a simulated population which matches the distribution of the target. The stochastic SA method on the other hand, as seen in Figure 5.3 does not yield the distribution of the target, instead it better replicated the seed (the NRTS) used in the spatial microsimulation.

The 8 variables used for the spatial microsimulation have on average (or typically) about 4 variable-categories each, making the modal density near the ~0.03 proportion (i.e. 1/(8*4)). Majority of the variables (other than the 'travel purpose' variable) do not present particular disparities in the population proportions simulated for the different variable categories within them, as such the mode of the density plot is located near the ~0.03 (i.e. 1/(8*4)). The software script used to generate the density plots are described in Appendix B.1, and included in the CD accompanying the thesis.

The results presented in Figure 5.3 illustrate the simulated population distributions, when the Census and NRTS data are combined: In this first spatial microsimulation, the NRTS (DfT, 2013a) is the seed which forms a skewed sub-set (not a random representation) of the target 2011 Census (Stillwell and Duke-Williams, 2003, UK-Data-Service, 2011, UKDS, 2016) flow data. The variables included in the NRTS seed are ‘Individual ID’, as well as those indicated in Figure 3.1. The Table 5.2 highlights these variables and that table on the left (within Table 5.2) identifies the variable categories adopted in the first spatial microsimulation procedure.

Table 5.2 The variables in the NRTS seed and Census target, as well as the colour-coded variable categories created.



1	National Rail Travel Survey – variables	No . of cat .
1	Purpose of journey	13
2	How often do you make this journey	6
3	Where did you come from	13
4	Postcode and address of initial origin	1
5	Postcode and address of final destination	1
6	Gender	2
7	Age	9
8	Cars or vans in household	4
9	People in household	2
10	Ethnicity	16
11	Income of household	7

2	2011 Census Interaction - variables	No. of cat.
1	Method of travel to work	12
2	Residence	1
3	Work	1
4	Gender	2
5	Age	6
6	Cars or vans in household	4
7	Family status	10
8	Country of birth	5
9	Social grade	5
10	Economic activity	5
11	NS-Sec	16
12	Industry	22
13	Occupation	10

Variable	Categories Created	No. of Cate-gories
Travel Purpose	Non-Train Commute (NTC), Train commute (TC)	2
Income	Below £7000 (Step_10), £7000-£12500 (Step_20), £12501-17500 (Step_30), £17501-35000 (Step_40), £35001-50000 (Step_50), £50001-75000 (Step_60), £75000+ (Step_70) as at year 2004/5	6
Age	Between 16-24yrs (ag16_24), 25-34yrs (ag25_34), 35-64yrs (ag35_64), over 65yrs (ag65_)	4
Gender	Male, Female	2
Household Cars & Vans	No car (Car_0), one car (Car_1), two cars (Car_2), three cars and over (Car_3),	4
Type of household	Household Type A, B, C, D, E (composed from the household people and status)	5
Ethnicity	European, Non-European	2

Note the colour codes used to identify the associated variables and categories. The origin and destinations (OD's) have not been included as categories as they are appended to the listed categories thereby creating the re-structured seed data.

The target Census flow data are already presented as $3D$ aggregates arrays made up of tables of location of usual residence and place of work by a range of socio-demographic attributes (including age, gender, income, ethnicity, commute mode, household-type, -cars and children). To facilitate reconciling the NRTS seed with the full conditional distribution of the Census target, the NRTS seed data is restructured. The restructuring creates a cross-classification of each individual variable category such that for example, an individual in the NRTS seed in the age bracket Ag16-24 who travelled from a residence in Postcode Sector LS2 5 to a work destination at BD3 4, would be reclassified as [LS2 5][Ag16-24][BD3 4]. This new reclassification would form a seed category as opposed to the previous category which was just [Ag16-24]. In essence the socio-demographic attribute categories within the NRTS seed are associated with an origin and destination, forming a more granular seed dataset. By so doing ~1351 seed variable categories and 2192 Census variable categories were established for spatial microsimulation.

The role of the Census interaction data is that of a set of $3D$ target tables, and as described earlier in Chapter 4, the m-IPF methodology makes provision for the inclusion of a combination of sets of different dimensioned ($1D, 2D, 3D, \dots nD,$ with $1 \leq n \leq \infty$) target arrays in the microsimulation process. The process after re-structuring the seed, involves identifying the matches between the seed and target variable categories. This would give and indication of the extent of sampling zeroes (as opposed to structural zeros). A decision is then made via a sensitivity analysis to decide the nominal amount required to be added to the seed table to alleviate the effect of structural zeros. The sampling zeros are accommodated since the m-IPF would simply yield a non-complete volume simulated population. The SA method would yield a requisite volume, albeit in many instances sub-optimal. The final results review aims to ascertain that the impact of adding values to structural zeroes has nominal effect on the results by yielding a simulated population with individuals having a combination of attributes which form a subset of individual attributes within the NRTS seed.

The results show that the deterministic iterative proportional fitting (IPF) spatial microsimulation strategy produced a synthetic population with similar distribution (and vertical median line) to the target Census. The stochastic SA strategy on the other hand created a simulated population distribution (and median line) similar to the NRTS sample seed, indicative that the proposal distribution from a skewed seed does not evolve to the target

Census distribution in the stochastic SA[43] spatial microsimulation. Under conditions where the seed sample is skewed or biased relative to the target, Figure 5.3 is a crucial result, and indicates that the deterministic m-IPF methodologies with multi-level simultaneous constraints applied are better at replicating the distribution of a target population than the stochastic MCMC based alternative. The reason and mechanism for this is that deterministic methods specify a proposal population from the seed, but with the distribution of the target data. This proposal distribution is iteratively improved subject to the objective function, until convergence to a target distribution. The ability in this instance to exploit the full variability in the target data, as well as an ability inherent in m-IPF to select fractional individuals ensures that non-absorbing regular iterative transitions are created (Bajram Spaseski and Ler, 2013) at each iteration step, ensuring convergence to the stationary target distribution.

With stochastic methods, statistically, repeated sampling of the seed will yield a distribution similar to the seed. Otherwise, subject to the constraints on the sampling process and no hill climbing element, repeated sampling will yield a distribution in which the constrained interacting variables follow the constraint. It is envisaged that the yielded outcome distribution (i.e. either similar to the seed or the constraint) would be conditional upon the skew between the seed and constrained variables. If the seed has dominance in this interaction, the simulation process will converge to the distribution of the seed and not to that of the intended target distribution, as seen in the SA method applied to the NRTS and Census. Further, the exploratory stage of the SA which involves sampling whole numbers of individuals[44], precludes fine scale optimization of the proposal distribution, contributing to the so-called 'sub-optimal trap' (Williamson et al., 1998). The stochastic sampling

[43] It is note worthy to emphasise that the SA approach implemented in the spatial microsimulation of the Census and the NRTS data have exactly the same set of cross-classified seed and benchmark constraint as the m- IPF strategy. The re-structuring of the seed ensures that the full conditional distribution of the constraint is the target.

[44] In the IPF procedure fractions of individuals are achievable. Recall that the IPF maps a seed population onto the target population such that, if the seed has say 4 individuals of specific required attribute, and the target population is 5, then each of the seed has a weight of 1.25. This indicates that a whole individual and a quarter fraction of an individual is simulated in the IPF. In the SA methodology on the other hand, only whole number multiples of individuals are created.

procedures are akin to the Metropolis Hasting algorithm which is premised on the proposal distribution being a good approximation of the joint posterior target (Lunn et al., 2012), hence the present limitation of the SA procedure.

The results in Figure 5.3 indicate that although the stochastic methods may yield commensurately lower TAE values, they do not yield the crucial distribution of the target data, especially in the instance where the seed is skewed relative to the target. This highlights areas of potential improvement of existing stochastic microsimulation strategies. In the stochastic algorithm despite the restructuring of the seed, there is no inherent feature of the methodology that requires that the simulated distribution and the target distribution are similar. Instead, all that is required in the SA method is that the absolute magnitude of the difference in volumes of simulated and target quantities be kept to a minimum. In the developed m-IPF method on the other hand, whilst the objective function[45] (Bešinović et al., 2013, Everitt, 2012, Rao and Rao, 2009, Törn and Zilinskas, 1989) is also the TAE, the re-structuring of the seed inherently also ensures that the fit to the target distribution is the objective function.

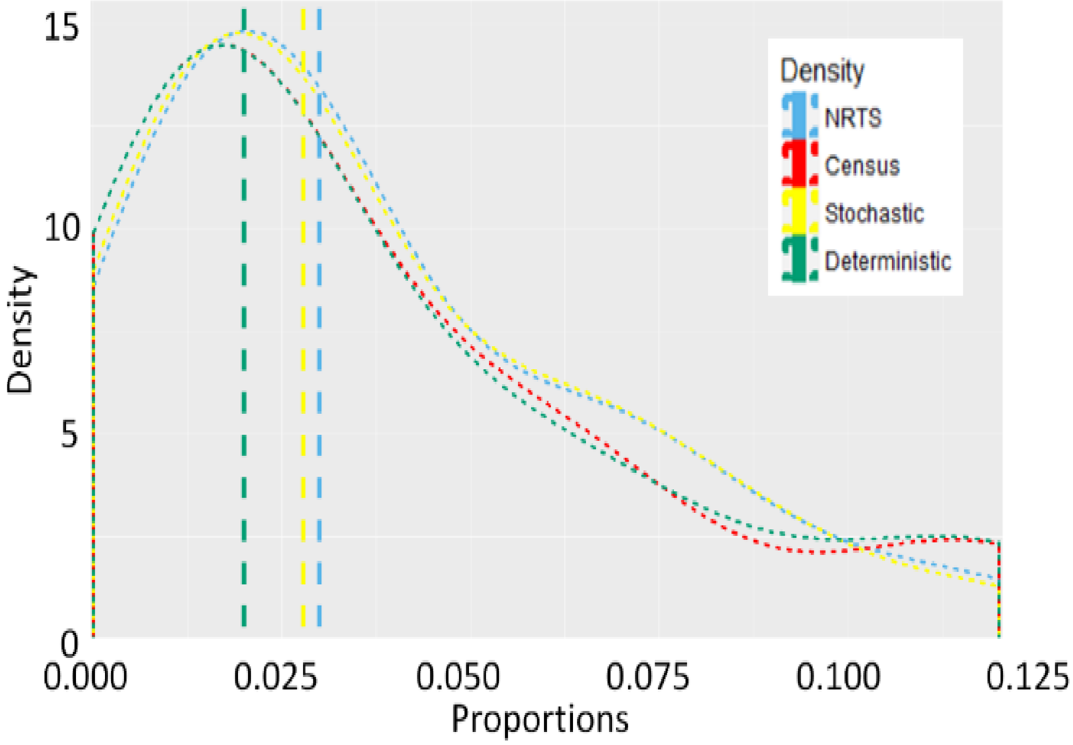


Figure 5.3 Distribution of Census, NRTS, m-IPF, SA populations.

[45] The objective function is the function that it is desired or required to minimize or maximize. In traditional spatial microsimulation it is desired to minimize the total absolute error (TAE)..

5.3.2 Stage 2 - Linking-in the LENNON ticketing data

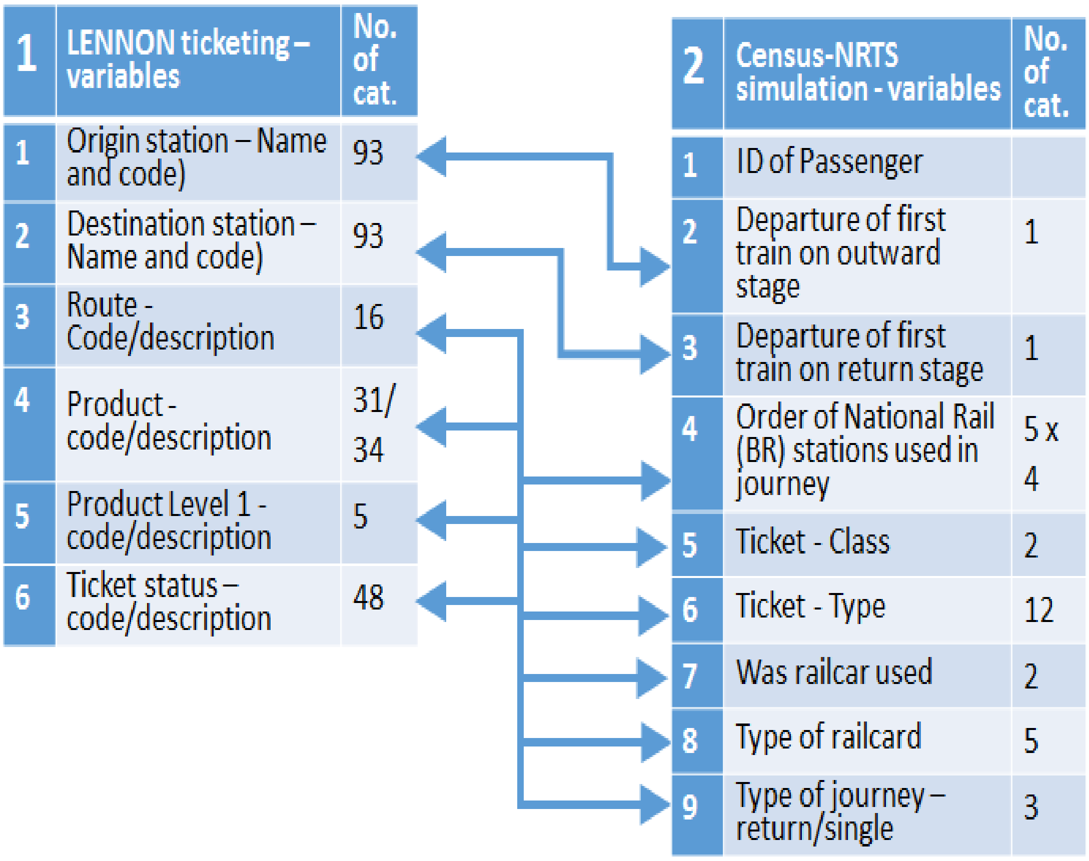
Once the Census interaction data and NRTS are combined to create a synthetic population, the challenge then lies in linking-in the LENNON ticketing data. As presented in Section 3.1, the LENNON ticketing data can be assumed to be a time series cross-section (TS-CS) dataset with each time period made up of 4 weeks. This assumption is made because the ticketing data available to the project had been aggregated into 4-week periods for confidentiality and anonymization. The end of each period coincides with rail industry setting rounds, such that substantial changes are not made to the rail service provisions in-between the setting rounds. On this basis the rail mobility dynamics can be assumed to be in equilibrium with only nominal changes during each period. As such, ticket sales in a period which belong to the same cross-section are iid (independent and identically distributed). For analysis, a cross-section of LENNON ticketing data for a 4-week period is extracted, forming the target dataset. The simulated population from the NRST-Census spatial microsimulation are then used as the seed 'sample' (albeit comprehensive).

The previous section (Stage 1 spatial microsimulation) demonstrated that if the full distribution of the target is used instead of the marginal), the m-IPF outperforms the SA by producing a better synthetic population having the distribution of the target. As such the m-IPF deterministic spatial microsimulation strategy has been adopted as the choice strategy for combining datasets when the seed are skewed and does not have the same distribution as the target data. When the seed distribution (in this case the simulated NRTS-Census micro-population) is a skewed representation of the target population (in this case the LENNON ticketing data representative of only the railways population), meaningful features of the joint probability distribution can be predicted better using the 'median' of the distribution, instead of the traditionally used 'mean', as a best guess of estimates of the average of the distribution[46] of the simulated population (Pearl et al., 2016). On these bases the median line as well as the curvature of the density plots is used to assess the similarity in joint distributions of the simulated and target populations.

[46] The choice of median minimizes the total absolute error, whilst the mean minimises the total square error. The literature however provides that the median line provides a statistically better guess of estimates of the variable of the distribution in complex scenarios like mobility interaction.

The simulated population with the full set of attributes from Census-NRTS data are combined with the LENNON ticketing data based on the shared attributes between these datasets. The shared attributes previously shown in Figure 3.1 include: station entry and exit, class of ticket, as well as variables relating to ticket, journey and passenger type. These variables are summarised in Table 5.3. As the simulated population from Census-NRTS already consisted of a cross-tabulation of the requite variables, only the LENNON data needed to be re-structured. The re-structuring serves two purposes: first it enables the assertion that the variable categories in the Census-NRTS data match those in the LENNON. Secondly, the re-structuring enables the origin-destination of each ticket recorded within LENNON to be attached to the other endogenous ticket attributes, thereby creating a $3D$ arrays and ensuring that the joint distributions of the seed and target are reconciled (as opposed to just the marginal).

Table 5.3 The variables in the LENNON ticketing data and those inherited from the NRTS during the Census-NRTS simulation.



1	LENNON ticketing – variables	No. of cat.
1	Origin station – Name and code)	93
2	Destination station – Name and code)	93
3	Route - Code/description	16
4	Product - code/description	31/ 34
5	Product Level 1 - code/description	5
6	Ticket status – code/description	48

2	Census-NRTS simulation - variables	No. of cat.
1	ID of Passenger	
2	Departure of first train on outward stage	1
3	Departure of first train on return stage	1
4	Order of National Rail (BR) stations used in journey	5 x 4
5	Ticket - Class	2
6	Ticket - Type	12
7	Was railcar used	2
8	Type of railcard	5
9	Type of journey – return/single	3

The pre-processing steps for reconciling the two tables are described in Appendix B.2. and the script included in the CD accompanying this thesis.

As seen from Table 5.3, the seed (Census-NRTS) in this second simulation consists of attributes inherited from the NRTS which are relatable to the LENNON attributes. In order to limit the clustering of the simulated population created, only $3D$ arrays in the seed are exploited at a time (i.e. the origin-destination and endogenous railway/ticket attribute. These are matched with the corresponding re-structured LENNON data which is constituted from the origin-destination and a ticket attribute. The R-script used to fulfil the second m-IPF are described in Appendix B.2, and included in the CD accompanying the thesis.

The results shown in Figure 5.4 show the distribution of the simulated NRTS-Census-LENNON micro-mobility simulated population, when the simulated (NRTS-Census combined with LENNON) population is sampled with probability distribution equal to the weights derived from spatial microsimulation. The distribution of the Census population has also been included as the second plot in Figure 5.4, to enable comparison with the simulated population. As seen in the density a plot of Figure 5.4, the median line of the Census is distinct from population derived using the m-IPF simulation weights. This is because the target distribution in this second simulation is the LENNON ticketing data which is distinct from the Census (even when a comparable set of variables are investigated).

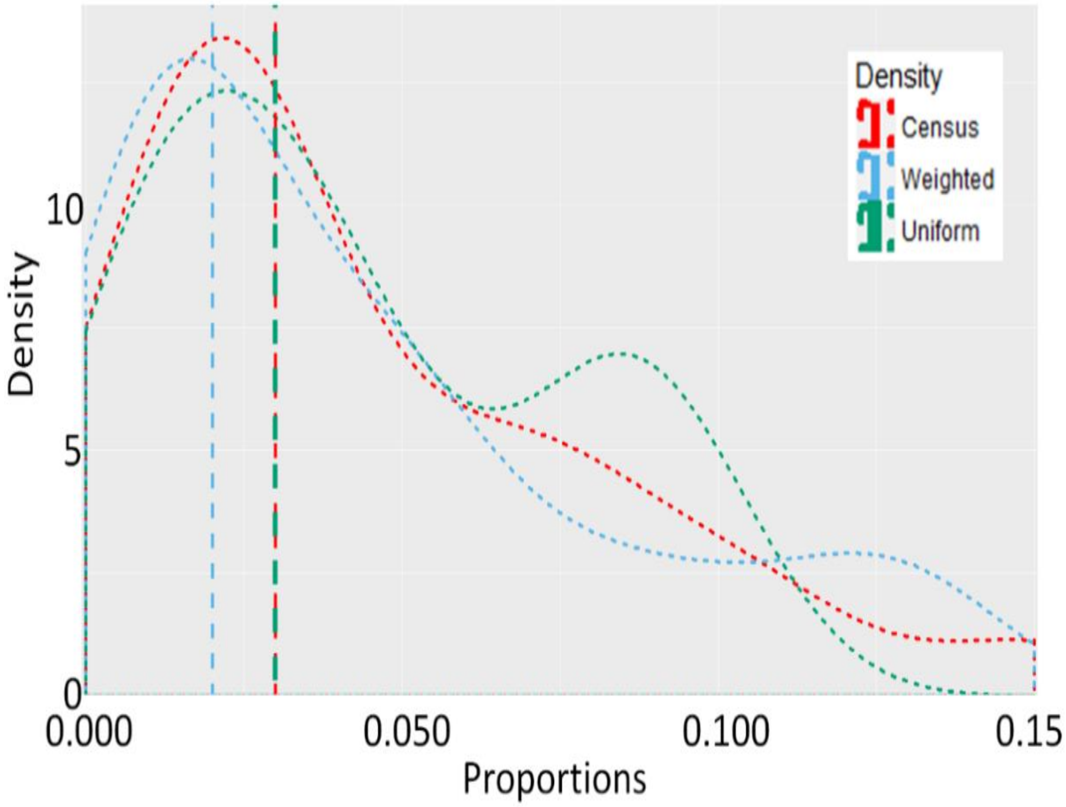


Figure 5.4 Distribution of Census, m-IPF (weighted), m-IPF (uniform).

The results produced validate the spatial microsimulation methodology as the simulated population replicates the target distribution in the proviso that the full distribution of the target is used as the constraint (and not just the marginal). In this instance, the stage 2 spatial microsimulation replicates the railways population according to the weights deduced from using the m-IPF spatial microsimulation strategy. An objective rich-attribute representative micro-level population of rail passengers are created. The micro-simulated population provides a cross-tabulation of the variables that facilitates relating population mobility to demographic characteristics, relative location, time of the day, and other likely drivers of behaviour and interaction.

The plot in Figure 5.5 supports the results produced from the stage 2 spatial microsimulation which combined the NRTS-Census with the LENNON data. The results (see Figure 5.5) reveal the distribution of the simulated individual level NRTS-Census-LENNON population when sampled with probability equal to the spatial microsimulation weights. The peaks and troughs of each plot in Figure 5.5 measures the proportion of individuals that fall within each variable category (marked by the horizontal black stripes). Observation of the plots reveals that the red (NTRS) and blue (NRTS-Census-LENNON) lines have similar troughs and peaks for each variable category displayed. This is especially so for those variables that are known in literature to be important attributes of rail passengers, for example travel purpose which in this instance are categorized into non-train commuters (NTC) and train-commuters (TC). This is emphasized in the inset of Figure 5.5.

To understand the plot in Figure 5.5, all the variables involved in the spatial microsimulation contribute a total unit value on the vertical axis. As such, if there are 6 variables in the spatial micrsimulation, then each variable will have total value of about 1/6 (approximately 0.17). Then, if the variable has say 4 categories, the value will be spread across the 4. Assuming that each of these variable categories have equal contribution to the simulated population, then the value for these variable categories on the vertical axis in Figure 5.5 would be about 0.04 $((1 \div 6) \div 4)$. If as seen in Figure 5.5 a particular variable, say travel purpose is made up of two categories TC and NTC (i.e. train commute, and non-train commute), and if one of the categories yielded a higher proportion of the simulated population, then in Figure 5.5 the prevailing category will be about 0.13while the auxiliary category would be much lower at say 0.04.This is the case for the population simulated from the Census-NRTS-LENNON data. This makes intuitive sense as train commuters would be the predominant simulated population.

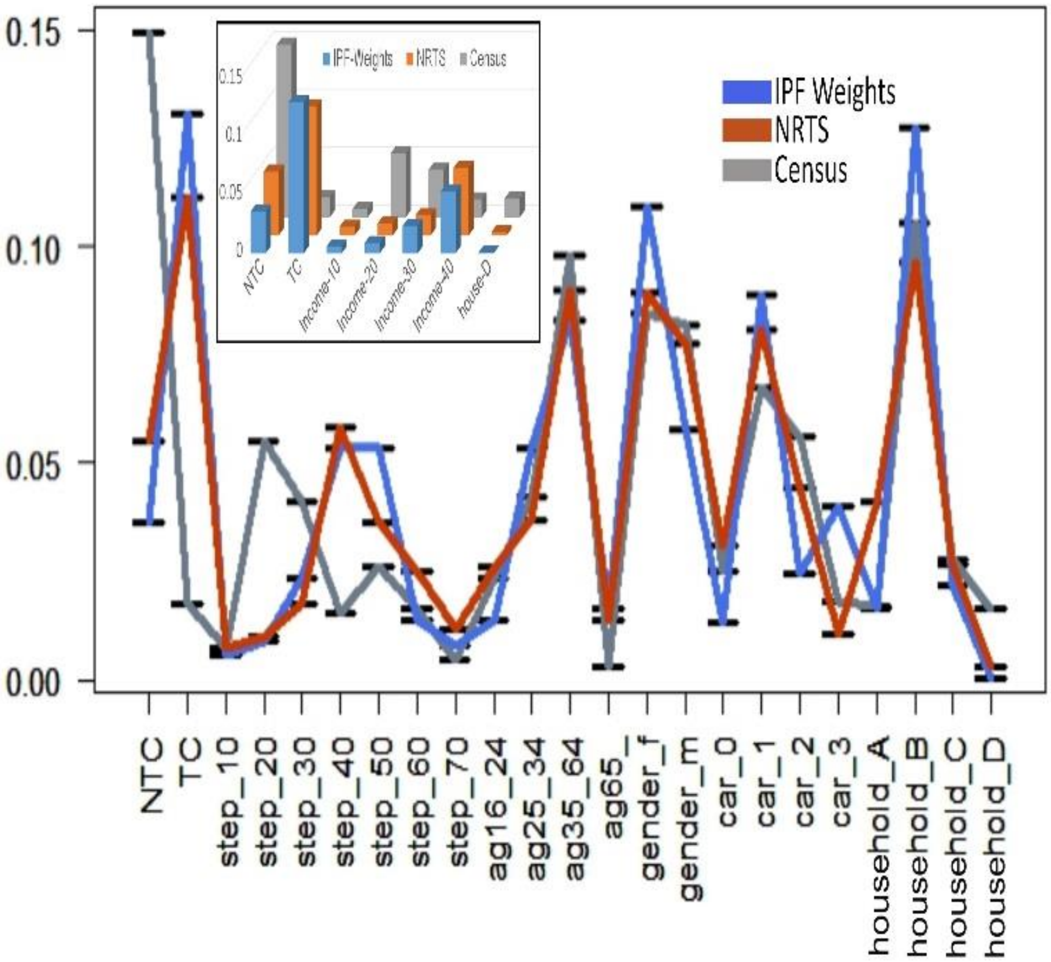


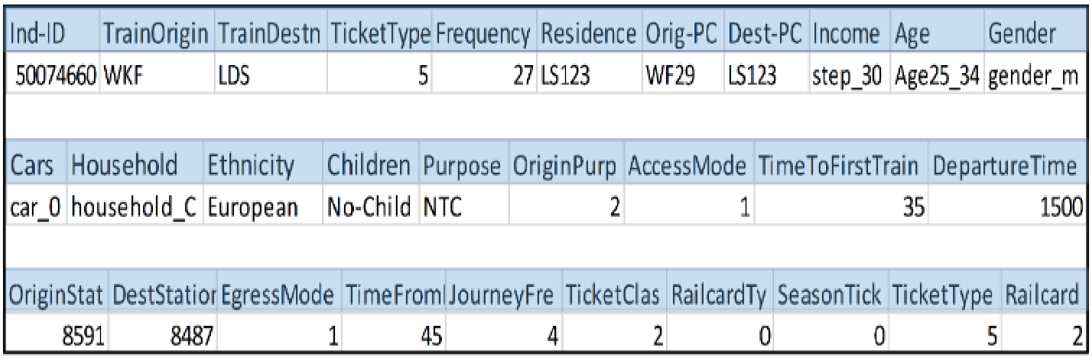
Figure 5.5 Variable categories for Census, NRTS, m-IPF.

The NRTS data (red plot) shows that whilst passengers who do not commute by train (NTC) form a 0.05 proportion (i.e. ~0.05*6 variables*100% = 30%), the passengers who commute by train (TC) form 0.11 (~0.11*6 variables*100%) = 66%. Similarly, the simulated population from the NRTS-Census-LENNON also shows NTC forms 0.04 (24%) whilst TC is ~0.13 (76%). Both datasets indicate that the proportion of passengers commuting by train (TC) is at least twice the non-commuters. This is intuitively anticipated of a survey constrained to railway passengers (the NRTS) and a representative simulated population of railway passengers (from combining NRTS-Census-LENNON). This is indicative that the individual-level population derived by IPF is representative of railway passengers. The much higher values of NTC than TC in the grey (Census) plot in Figure 5.5 is quite the opposite of the red and blue plots, and reflective of the distribution in the wider Census population where a majority of passengers do not commute by train (NTC). Typically only ~8% (Stillwell, 2006) of the wider UK population commute by train (TC).

5.3.3 Cross-tabulations of variables

As a validation of the population synthesis data, and to show the level of detail achievable from the m-IPF hierarchical constraint methodology, a snippet of micro-level cross-tabulation results are shown in Table 5.4. The result shown is for passenger ID-50074660. This passenger is described as NTC, indicating non-commute to work by train (however not precluding the occasional use of the train for other purposes). The time of arrival at the station is 35mins before the first departure train at 15:00 pm. The passenger originally came from Postcode Sector WK29 and accessed the train service at Wakefield Westgate station (WKF). The passenger's final destination, in this instance the residence is in Postcode Sector LS123.

Table 5.4 Query of m-IPF cross-tabulated passenger attributes.



Ind-ID	TrainOrigin	TrainDestn	TicketType	Frequency	Residence	Orig-PC	Dest-PC	Income	Age	Gender
50074660	WKF	LDS	5	27	LS123	WF29	LS123	step_30	Age25_34	gender_m

Cars	Household	Ethnicity	Children	Purpose	OriginPurp	AccessMode	TimeToFirstTrain	DepartureTime
car_0	household_C	European	No-Child	NTC	2	1	35	1500

OriginStat	DestStatior	EgressMode	TimeFroml	JourneyFre	TicketClas	RailcardTy	SeasonTick	TicketType	Railcard
8591	8487	1	45	4	2	0	0	5	2

Further visually mapped query results are shown in Figure 5.6 and Figure 5.7. The results are intuitive and this provides some external validation of the micro-simulated population. Figure 5.6 are results of a query of the population who reside in Postcode Sectors (PS's) in Bradford, who commute to work by rail, have a Rail card, regularly buy a return ticket, travel within 15miles of their typical residence, earn between £17.5 – £35k (at 2011 rates) per annum, and live in a household with no car and no children. The results indicate higher volumes of passengers who are more likely to use the train based on their proximity to a rail station (shown by the green dots with the connecting curves as rail lines). Figure 5.7 shows passengers who reside in Leeds (LS), who do not commute to work by rail (NTC), and live in a household with three or more cars (Car-3+). It is seen that PS's in the vicinity of the railways stations, tend to have a higher number of passengers as they have easier access to the rail network. The high volume of passengers simulated for the LS179 Sector (near the middle of the map in Figure 5.7) are perhaps reflective of an affluent neighbourhood, occupied by households with 3 or more cars (Car-3+), and who do not commute by train (NTC).

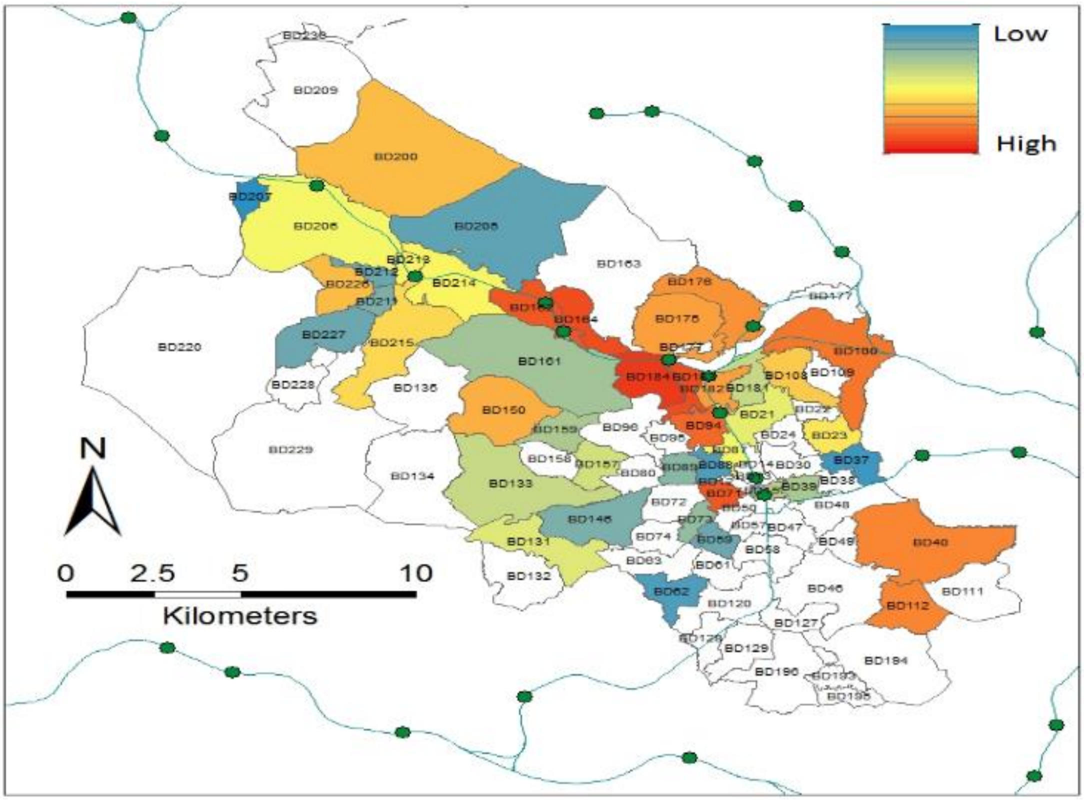


Figure 5.6 Query results for Bradford (BD)[47].

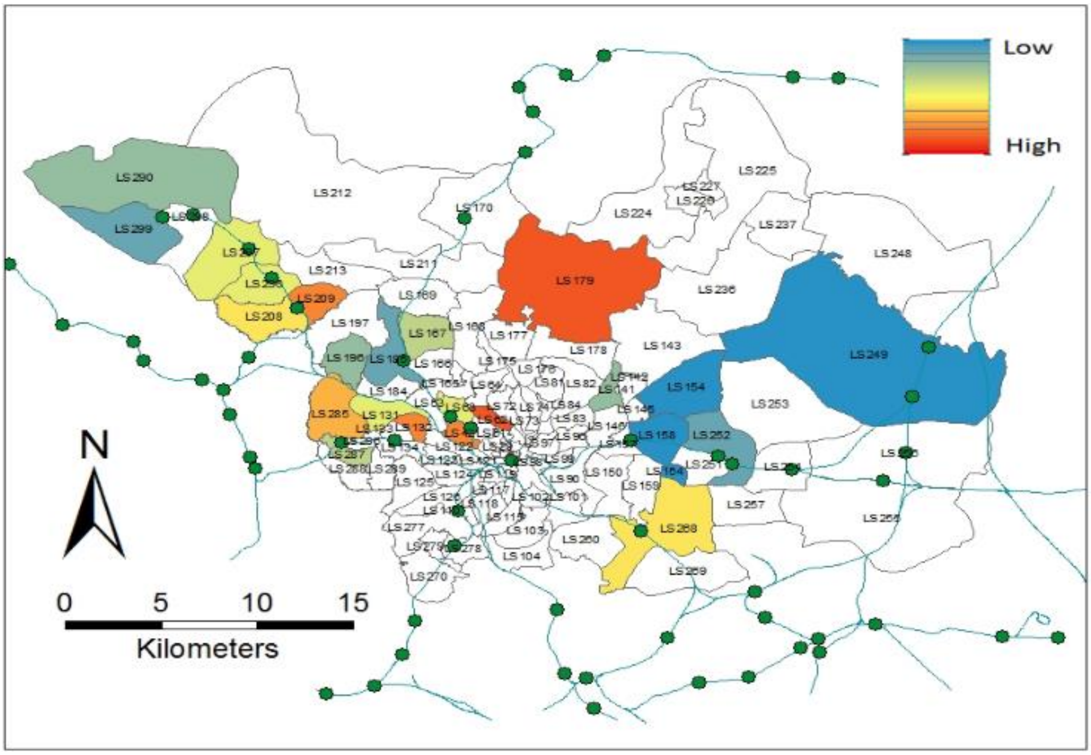


Figure 5.7 Query results for Leeds (LS).

[47] The white zones in Figure 5.6 and Figure 5.7 indicate the zones where a proportion of less that 5% of the population satisfy true to the query.

5.4 Discussions and conclusions

Spatial microsimulation enabled the combination of the various attributes in the disparate datasets, and produced granular cross-tabulations of the combined variables in the datasets. Such granularity and cross-variability in individual passenger attributes can only be achieved through spatial microsimulation. Spatial microsimulation also enables the independent socio-demographic and within network variables defined in the Census, NRTS and LENNON databases to be dependently cross-tabulated, showing the previously unavailable variability of passenger attributes at Postcode Sector levels.

The combined results from Chapter 4 and this Chapter 5 show that the stochastic methods (SA) are less sensitive to changes in seed sample ratios and skew, requiring a marked change in sample ratio to affect the stochastic TAE values. An increase in the number of constraints yields lower TAE values for both the deterministic and stochastic strategies. The standard deterministic IPF shows an improvement in the accuracy of the TAE results as the sample ratio is increased and as the skew in the seed data is reduced. The standard IPF however produces poorer TAE values than the stochastic SA. In this chapter, the standard IPF was developed by incorporating multi-level simultaneous hierarchies in the seed, thereby enabling the exploitation of the full joint distributions in the target. When this new methodology was applied to the test case of mobility in the West Yorkshire railways, the deterministic m-IPF method yielded a simulated population with a distribution similar to that of the target population, whilst the stochastic method produces a simulated population distribution similar to that of the seed data. This warrants the preference for the improved m-IPF method in replicating a target distribution, and in being better suited for spatial microsimulation when the seed data is a skewed representation of the target population.

This chapter has applied the deterministic m-IPF procedure to combine data from NRTS, Census and LENNON to create an attribute-rich representative individual-level population of railway passengers. The spatial microsimulation case study in this chapter has been presented in a practice oriented accessible way to highlight the mechanisms of the different spatial microsimulation strategies, and to enable the results to be reproduced. The potential deficiencies in the classic IPF when applied to skewed consumer big data have as such been addressed by developing a structured multi-level iterative proportional fitting (m-IPF) algorithm.

The stochastic methods are variants of the Markov chain Monte Carlo (MCMC) algorithm. The HC and SA spatial microsimulation strategies are variants of the Metropolis-Hastings MCMC algorithm. The implementation of the HC and SA strategies are based on the TAE being the objective function. An optimisation therein guarantees a low TAE, but this does not guarantee that the simulated population created would have the distribution of the target. To yield the target distribution using the SA and HC methods, a measure of the difference in distributions of the seed and target are required for inclusion as the objective function. To the best of our knowledge, such MCMC chains are currently not available in the literature. The simulation based population synthesis methods are variants of the Gibbs MCMC algorithm. In this chapter, our statistical formalisation shows that Gibbs sampler based population synthesis (Farooq and Bierlaire, 2017) is conceptually inconsistent in the current state of development. When the seed data is skewed relative to the target, the missing aspect of the seed cannot be described as missing at random relative to the target. Under such conditions, inference about the target cannot be based on just the seed data (Acuna and Rodriguez, 2004, Allison, 2001, Little and Rubin, 2014, Rubin, 1976, Schafer and Graham, 2002), which would be described as statistically missing-not-at-random (MNAR). A missing data model would as such need to be incorporated in the spatial microsimulation process. To the best of our knowledge, such Gibbs MCMC based imputation within population synthesis or spatial microsimulation are currently not available in the literature. These methods have promise in exploiting the strengths of MCMC methodologies, but their application at the current stage of development requires particular expertise on the part of the data analyst.

The internal and external validation of the spatial microsimulation results derived in this chapter is discussed below. References to the typical challenges particularly in external validation are also discussed. Then, particular precautions in practical implementation of the spatial microsimulation strategies developed in this chapter are presented. These precautions are pertinent in ensuring robust solutions in a wide range of application scenarios. The chapter ends by making conclusions about the current choice of spatial microsimulation strategy to adopt when reconciling skewed consumer data with other datasets from randomized control experiments or from comprehensive measured surveys of a wider population.

5.4.1 Validation of results

The MOIRA[48] is the rail industry's reference principal planning tool, including a comprehensive representation of travel on the railways and used for planning by estimating flows and linking the service supply side of the rail facility to passenger demand. MOIRA is used to create the origin-destination matrix (ODM) which is a contingency table of flows between every train station in the UK, and it is considered a reference dataset. MOIRA data are typically released as estimates of annual flows on the rail network, thereby allowing a comparison with aggregated microsimulation results. The West Yorkshire ticketing data used for analysis consists of only those flows emanating from and terminating in the West Yorkshire County. In addition, tickets sold by regional transport executives (PTE's) have not been included. Flows associated with PTE tickets and restrictions to West Yorkshire (WY) flows are estimated to account for about 55% of all flows. For validation, the flow volumes estimated from the developed m-IPF spatial microsimulation are compared to MOIRA estimates. As the m-IPF simulated population are restricted to only those emanating and terminating in West Yorkshire, for comparison the MOIRA flows are processed to exclude all non-WY flows. As a result, the difference in MOIRA and m-IPF flows would be expected to be of an order of magnitude of regional PTE ticket flows.

As stated above, MOIRA flows are aggregated annually, but in addition MOIRA does not distinguish flows to specific stations within Group Stations[49]. The m-IPF on the other hand disaggregates flows to group stations. As such, for comparison of MOIRA and m-IPF, it has been found simpler to exclude those flows associated with group stations. Excluding flows with entry and exit stations outside WY, and excluding group station flows, annually MOIRA yielded 47m flows while m-IPF yielded 18m flows.

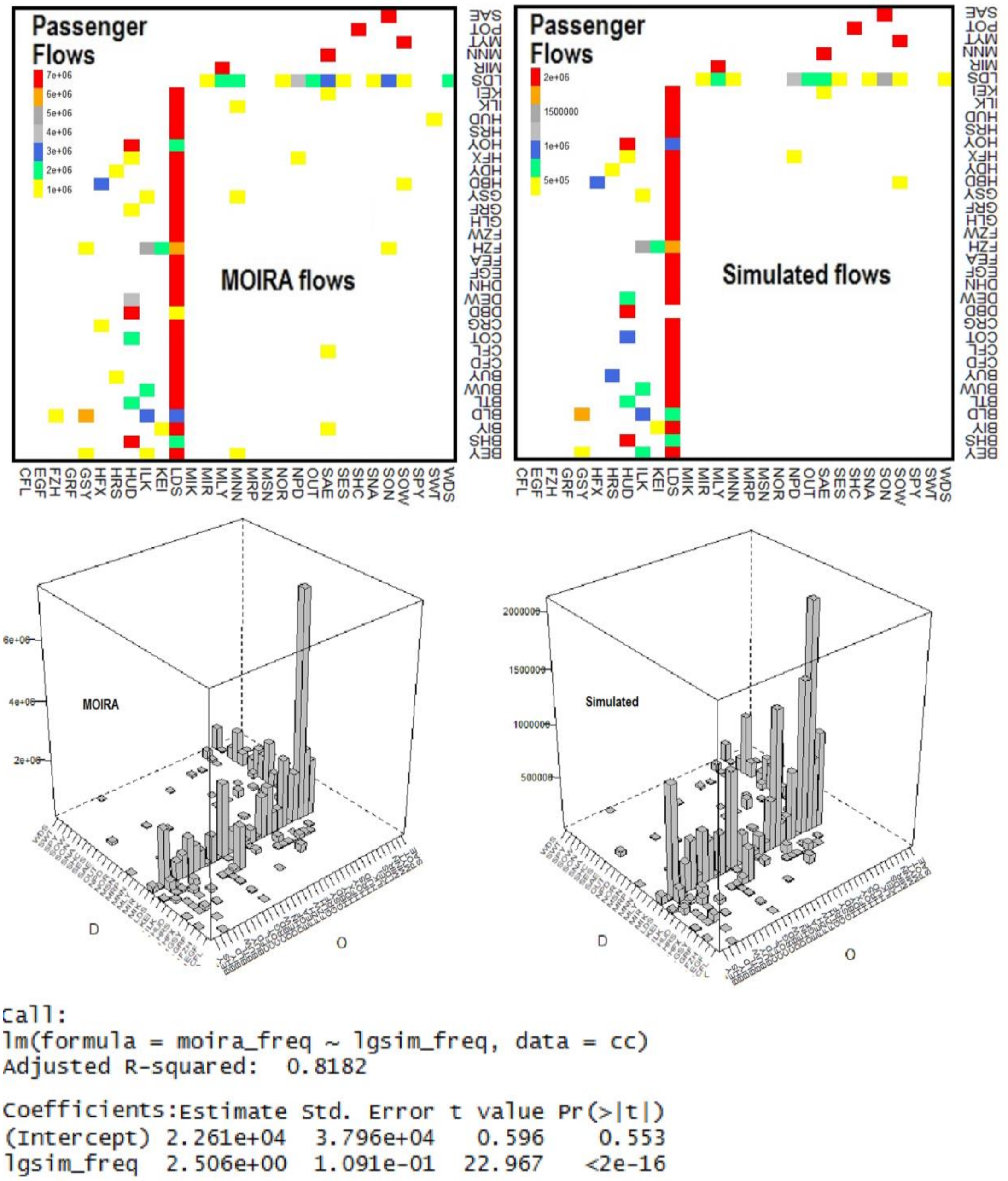
[48] MOIRA is the acronym for 'model of inter-regional activity' which is the railway industry reference demand forecasting software to assess impacts of supply service (time-table) changes on passenger demand, enabling strategic planning.

[49] British network rail (BR) tickets are normally issued to and from individual stations. When multiple stations with different access routes exists in a locality, BR has found it exigent to issue tickets which enable passengers choose any of the stations in the locality to access the train service. These BR tickets bear the name of the locality, without mention of any specific station in the locality. Such local stations are classed as group stations.

The comparative results from MOIRA and the simulated population from m-IPF are shown by the heat maps and the 3D bar charts in Figure 5.8. The heat maps are chosen to highlight the similarity in the trend of results derived from MOIRA and m-IPF. The left hand plot is the MOIRA plot and notice the large volumes recorded for flows associated with Leeds train station. The m-IPF plot on the right of Figure 5.8 was created by querying the m-IPF simulated population to aggregate flows by origin and destination stations. Despite the station-by-station similarity in the trend of MOIRA and m-IPF results, the m-IPF estimates are typically of the order of less than half the size of MOIRA flows. This is revealed in the 3D bar charts in the middle of Figure 5.8, where the left 3D bar chart represent MOIRA flows whilst the right are m-IPF flows. Note the scales of two bar charts in Figure 5.8 which indicate that the predicted m-IPF flows for Leeds station are about half the MOIRA flows for Leeds. This trend continues for the large volume stations within high population zones where there would be greater confidence in the results of the spatial microsimulation.

The periphery stations (see Figure 4.6, Figure 5.1, and Figure 8.1) where only a part of the Postcode Area geography falls within the West Yorkshire study area, reveal large and varied disparity between the MOIRA and m-IPF estimates. There could be an explanation for the large variance in MOIRA and m-IPF results for small volume and peripheral stations in the West Yorkshire (WY) study area. Non-publicly available MOIRA has known accuracy issues (Taylor, 2013b, Wardman et al., 2007), and m-IPF would yield a reduced population anytime the seed does not fully represent the content of the population. In addition, with m-IPF, all peripheral Postcode Areas that intersect the boundary of WY are clipped. The consequence of this is that the population simulated for such clipped zones, are equal to the population of the whole zone multiplied by the proportion of the clipped zone. If an Postcode Area (PA) has a total population of say 120, and if two-thirds (2/3) of that PA falls outside the WY study area boundary, then the maximum population the m-IPF could simulate for PA would be only 40 (i.e. the pro-rata proportion of the population occupying PA). On the other hand, MOIRA in such a situation would yield the full volume of the PA because passengers that fall on the part of the PA outside WY can still access the train station within the PA. This feature would add to the discrepancy between m-IPF and MOIRA flows. It is observed that peripheral stations in the WY study area have markedly lower volumes when MOIRA flows are compared to m-IPF flows. This scenario might be further exasperated when the flows associated with such peripheral stations are low compared to the regional average.

A regression of the MOIRA and m-IPF station-to-station flows as indicated in the text at the bottom of Figure 5.8; show a significant R^2 value of about 82%, indicative of a similarity in the trend of station-to-station flows between MOIRA and m-IPF. The parameter of regression indicates a value of 2.5, indicative that the m-IPF flows are about 40%of MOIRA flows. The non-perfect R^2 value (<100%) is reflective of those stations where there are disparities in the MOIRA and m-IPF results. These flows are represented by the yellow squares seen dotted around the middle right, lower right, and parts of the lower left extreme of the OD heat map in Figure 5.8.



```
Call:
lm(formula = moira_freq ~ lgsim_freq, data = cc)
Adjusted R-squared:  0.8182

Coefficients:Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.261e+04  3.796e+04   0.596    0.553
lgsim_freq  2.506e+00  1.091e-01  22.967   <2e-16
```

Figure 5.8 Comparison of MOIRA and simulated m-IPF station flows.

The external validation adopted has been to infer that the query of the cross-tabulated synthetic population yields results that are intuitive, and expected of mobility phenomena associated with the Postcode Sectors being queried. External validation has been reported to be more difficult to achieve (Lovelace and Dumont, 2016), and in the case researched the only reference data was the aggregate data from the MOIRA. Availability of smart card and OD surveys data from WYCA-ITA in the future would provide an opportunity for externally validation. Internal validation of the spatial simulation results is achieved using the total absolute error (TAE) convergence criteria, which is satisfied prior to terminating the microsimulation procedures. In addition to this, a visual inspection of the density distribution of the simulated population variable categories (as shown in Figure 5.3 and Figure 5.4) would reveal if the simulated population has the appropriate distribution of the target population. Further to this, the median line is a good indication of the similarity in the distributions of the seed data and the target distribution.

5.4.2 Precautions and practice

The inclusion of the identity (ID) for the about 50,000 seed samples for the West Yorkshire County meant an additional dimension of 50,000 factors (or characters) were added to the seed array, commensurately increasing the computer RAM requirements. The project programs developed using R-software required a maximum of 350GB computer memory and usable storage (RAM), but the upper limit of R data object capacity was restricted to 2^32 (5GB) by the R-studio platform. The choice of variables and data tables were managed to fit this 5GB limit, and the Arc3[50] computer (University-of-Leeds, 2017) was used.

Once the ID's are included in the stage 1 simulated population (NRTS-Census), it is possible to couple to the additional variables in LENNON as illustrated in Figure 3.1. By so doing individuals can be linked to the LENNON ticketing data. In many microsimulation strategies, the individual identities are typically not revealed, and the difficulties in estimating these are often excused by the fact that a concerted drive to develop these has not been pursued due to a need to maintain confidentiality or anonymization of data. Strategies that do not provide identity information for the simulated micro-population are typically called population reconstruction instead of spatial microsimulation (Birkin and Clarke, 1988).

[50] This supercomputer is 2-Linux CentOS(7) based, with 24 cores and 768GB of memory, with 350TB usable storage.

The code (R-scripts) used to implement the experiments on stochastic and deterministic spatial microsimulation strategies, using data from the NRTS, and then Stage 1 / Stage 2 spatial microsimulation, are respectively included in Appendices A.1-A.5, and Appendices B.1-B.2. These scripts have been developed from R packages ‘mipfp’ and ‘sms’ (Barthelemy et al., 2016, Kavroudakis, 2015, Team, 2000), and care has been taken to wrap the pertinent aspects of the scripts inside functions to make the script versatile for use in the range of constraint structures typical found in general spatial microsimulation problems.

In implementing the m-IPF procedure, the zero-inflated design matrices result in divisions by zero which crashes the algorithm. To forestall this, it was necessary when dealing with count data from mobility interaction (which inherently tend to be zero inflated), to add a nominal value to the zero entries in the data frame. Typically as recommended in literature (Lomax and Norman, 2016), the nominal addition would be of an order typical of the expected error margins in the spatial microsimulation. The effect of the nominal addition can be investigated by assessing the spurious cross-tabulated variability introduced into the simulated population. If the nominal addition is sufficiently small, this cross-variable micro-population would be negligible in the simulated population.

Another issue with simulated population at the current state-of-the-art is that as seed data are typically of low sample rations (≤5%), the tendency is that the simulated populations would consist ofclusters of individuals that would have has to be replicated many times to create the requisite population. The clustering is amplified in scenarios where the full conditional distribution of the target is exploited (as in the 3-level hierarchical constraint introduced in this research). The increasing availability of big-data for use as seed in spatial microsimulation has the potential to alleviate the clustering.

To curb the impact on accuracy of boundary areas that do not fall completely within the study area, the research would have to extend the study area substantially further that the area of interest in analysis. This would commensurately impact on the hardware and software requirements of a solution strategy. This is a particular area where current Bayesian learning strategies can be developed for spatial microsimulation, as these Bayesian strategies require commensurately less computer memory to run, but could be time-consuming in implementation.

5.4.3 Conclusions

The conclusion from this chapter is that the improved m-IPF deterministic strategy is preferred for use in spatial microsimulation over current stochastic strategies when the seed sample is a skewed subset (not a random representation) of the target population. A marked majority of revealed consumer data like the rail sector ticketing data, loyalty card data, social media data, and many surveys like the NRTS on the railways are measurements on a skewed subset of the wider population. These skewed datasets are best combined with other more representative datasets using the deterministic m-IPF procedures, provided also that multi-level seed data constraint are developed to enable reconciliation with the full joint (or conditional) distribution of the target constraint.

A novel contribution of the spatial microsimulation methodology presented in this research is in identifying the behaviour of deterministic and stochastic strategies when, as is typically with big data scenarios, the seed data is skewed relative to the target distribution. For such cases, the simultaneous hierarchical constraint methodology develop in this research identifies the superiority of the m-IPF strategy over MCMC based Metropolis-Hasting population synthesis. The MCMC based Gibbs sampling strategies adopted in the literature are not conceptually consistent for application to skewed data without the inclusion of a missing data model. The mechanism through which the m-IPF strategy converges to a stationary distribution of a desired target population when the seed is skewed is highlighted.

The methodology developed in this research explains the mechanism of the weights[51] derived from m-IPF strategies, and addresses the problem of ‘zero cell values’. The hierarchical constraint also incorporates mobility interaction in the population synthesis, thereby providing a more robust conceptual framework for the spatial microsimulation process. There are no papers in literature that have simultaneously used the origins and destinations of mobility interaction data as a basis for creating a synthetic population.

[51] The weights derived for each individual simulated in the m-IPF spatial microsimulation are indicative of two things: the first is how representative that individual is of a particular OD flow, and secondly the weights are indicative of how non-representative a particular individual is within the seed sample taken for a particular OD flow.

Chapter 6
GIS-GTFS NETWORK SIMULATION

In this chapter, the micro-mobility of passengers on a UK railways system is reproduced building a GIS network model logistically constrained by transit schedule information called the General Transit Feed Specification (GTFS), thereby creating a GIS-GTFS network model. The micro-level synthetic passengers created from spatial microsimulation, in the previous two chapters are used as input to the GIS-GTFS network model. The thinking is that if a representative passenger population is fed through a representative logistical rail network, the result would replicate each passenger's experience on the rail network. An overlay of each passengers experience and activity on the rail network would reveal the context of passenger mobility, revealing information like the crowding in the train when each passenger made their journey, the number of stops each passenger made, or the waiting times they might have endured. This framework enables the reconciliation of a simulated spatial located micro-level population outside the network with the within-network micro-mobility phenomena. The combination of socio-demographic and contextual information produces an attribute-rich population including both the exogenous and endogenous network attributes. To the best of our knowledge, no application exists in the research literature that utilizes GTFS information within a GIS, and uses this framework to harness passenger ticketing information. This chapter presents the GIS-GTFS network model in a practice oriented way, to enable the recreation of the results. The micro-level passenger information produced are potentially useful as input to rail sector planning models. The GIS-GTFS model is useful for management, maintenance, and interventions assessment on the railways.

The attribute-rich contextual information on mobility at individual-level derived from the GIS-GTFS model is subsequently used, as demonstrated in the next chapter (Chapter 7), as a basis for identifying passenger behaviour, like missing daily trip rates with season tickets, and precise flows to stations clustered into groups. Such additional behavioural information are necessary in order to identify particular drivers of mobility. In Chapter 7, Bayesian modelling is used to impute these behavioural attributes into the output from the GIS-GTFS model, and to simultaneously analyse and infer specific drivers of mobility phenomena on the railways network.

6.1 Introduction to transport system

In traditional transport modelling, say the classic 4-stages of trip generation, distribution, choice of mode and route assignment, the methodology is as follows: a projection of the number of trips from each region is made and a spatial interaction model (Batty and Mackie, 1972, Fotheringham, 1983) is used to sub-divide these flows to regions. A choice model decides which route passengers take, and then these are allocated iteratively to attain equilibrium in the system. The 4-stage model and variants of it have been the reference transport models and have been widely practised in the transport industry (de Dios Ortuzar and Willumsen, 1994, Hensher and Button, 2007). National country specific transport models (STAN, TEM, SMILE, SCENES, NTM) (De Jong et al., 2004, RAND-Corporation, 2017) have been developed, as well as more sector specific models tailored to different transport sectors (MOIRA, PDFH[52] etc.) (Worsley, 2012). The MOIRA is the reference transport model in the UK railway industry, while the PDFH is the reference demand forecasting tool.

The modelling framework developed in this thesis combines spatial microsimulation from geography, with a transport GIS network model to provide an alternative framework to the 4-stage model. Novel sources of large consumer data reflect passenger utility and spatial microsimulation enables attribute-rich micro-level demand populations to be synthesized. The recent availability of detailed transit schedule information enables the creation of logistical GIS network models. When the demand population are fed through a logistical GIS network through the use of real-time GTFS[53] information, estimates of passenger movement within the rail network can be achieved. This GIS-GTFS network dataset enables a query to identify all the micro-level within network (endogenous) passenger attributes. These attributes when combined with those derived from spatial microsimulation create an attribute-rich micro-level population, suited for detailed mobility

[52] The PDFH is the Passenger Demand Forecasting Handbook, is a core source of evidence and a reference strategy for forecasting railway passenger demand in the UK. The PDFH provides guidance on changes in passenger demand due to service quality, fares and external factors. on rail demand.

[53] Recall the GTFS was detailed in Section 3.1 and in the Class Diagram of Figure 3.2, and defines a common format for public transport schedules and the corresponding geocoded information.

analysis and spatial analysis. It is the availability of novel consumer data sources and GTFS information that has enabled the demand population and GIS network simulation (De Jong et al., 2007), with the potential to alter conventional transport modelling genres. The novelty of the GTFS information within a GIS network analysis, combined with spatial microsimulation developed for skewed big data, enables our method for generating passengers, distributing these passengers to choice railway stations, and allocating them to specific routes, more accurately than would have been possible using the traditional 4-stage methods (Zhao and Kockelman, 2002).

6.2 Methodology for GIS-GTFS network

A Geographical Information System (GIS) produces layered thematic visual results. The integrity of the layers is sustained by the strong data-typing structure of GIS which ensures that only similar specific information are contained within each layer. The strengths that GIS technology can bring, as such, include structured layered features, accurate and precise feature classes and attribute data (Odiari, 2011). At the core of the structure of GIS are data models which provide the relational database for coupling, querying and analysis of the disparate datasets, thereby making GIS readily suited to act as an integrator which synergizes the applications and tools interacting with it. These strengths of GIS facilitate topologically rich analytical models, enabling GIS to form computable physical models.

The GTFS[54] (as mentioned earlier) defines a standard geographic format for public transport schedules. The Class Diagram for GTFS information was introduced in Chapter 3, and detailed in Figure 3.2. A class diagram illustrates the relationship between the variables within the GTFS tables. A use case scenario (Adolph et al., 2002) can then illustrate how the GTFS information integrates the activities of the different operators, for example: standards developer, train operators, application developers and train customers. Such conceived use case scenario enables a relational data

[54] The GTFS information does not contain information on the number of train carriages in use on particular schedule route and trip times. The network operators have such information available, but this was not part of data supplied to the research project. As such, the passenger count at a location in time is used as a proxy for any potential crowding on the rail network and in the train carriage.

model (Codd, 1970) to be created for the GTFS information, enabling the development of a GIS to relate to, and query the GTFS. Integrating a GIS network dataset for the transport network with the GTFS information enables the solution of the route of passenger travel, constrained by the service provisions of the rail network. The GIS-GTFS model presents a scenario whereby an attribute-rich demand population inherently containing the utility maximizing behaviour of passengers can be assigned to the GIS-GTFS network traffic. This work as such has the potential to radically alter traditional multi-agent transport simulation genres by precluding the need for utility maximizing traffic assignment and the inherent complexities therein.

A GIS-GTFS model as such forms the computable physical model of the transportation network. In this research, the synthetic passengers from the spatial microsimulation (developed in Chapter 4 and Chapter 5) are used as input to the GIS-GTFS network model created in this research. Overlaying each passenger mobility reveals the space-time context of mobility and activity on the logistical rail network. The methodology for the GTFS use case scenario, the GIS network model and the combined GIS-GTFS network are presented in the next sub-sections.

6.2.1 GTFS use case scenario

The GTFS information on its own enables the identification of the precise location of the trains. However, when combined with a GIS network dataset, the location of train passengers in transit on the rail terminals, platforms and within the trains can be solved. The GTFS was originally developed as a lightweight Google standard (Antrim and Barbeau, 2013, McHugh, 2013), and consists of datasets from agencies on trips, routes, stops, and stop times, schedule calendar dates and exceptions, geographic trip shapes, transfers, service frequencies, and feed metadata (Google, 2016). The GTFS information from ATOC and DfT (TransitFeeds, 2017) are regularly updated to reflect the correct up-to-date transit activity, and published at time intervals ranging from between 5 to 9 days, at the end of each monthly ticketing period. The industry aim for producing the GTFS is to enable sharing of data and standards between the interested parties in rail sector. The functionality of the GTFS is captured in the Use Case diagram in Figure 6.1. The variables within the data tables of the GTFS information were represented by the Class Diagram shown in Figure 3.2 (Chapter 3, Section 3.1). The class diagram facilitates the creation of a GIS data model to reconcile with the GIS to simulate the precise location of every in-service train on the railways network within West Yorkshire.

The GTFS nomenclature defines trips, shapes and routes. A trip represents the journey made by the train in going through the stops in a time-specific sequence. The trips produce the shape traced out by the train. Routes specify the service lines provided by the network operators to enable mapping the network. In essence, the trips form the shapes which are combined in parts to form the routes. In the UK, the railway lines have evolved with advances in signalling to permit flow in both directions, unlike in many countries where the transit is only double tracked so that trips occur in only one direction along the transit lines. As such, in the GIS-GTFS developed for the UK railways, transit evaluators were specified to apply to both the to-transit and fro- transit directions. The ArcGIS created for the GIS-GTFS model required extensive coding input (running into a full year) as detailed in Appendix C, and was not a 'plug-and-play' model typical of say MATSim (Gao et al., 2010, Horni et al., 2016).

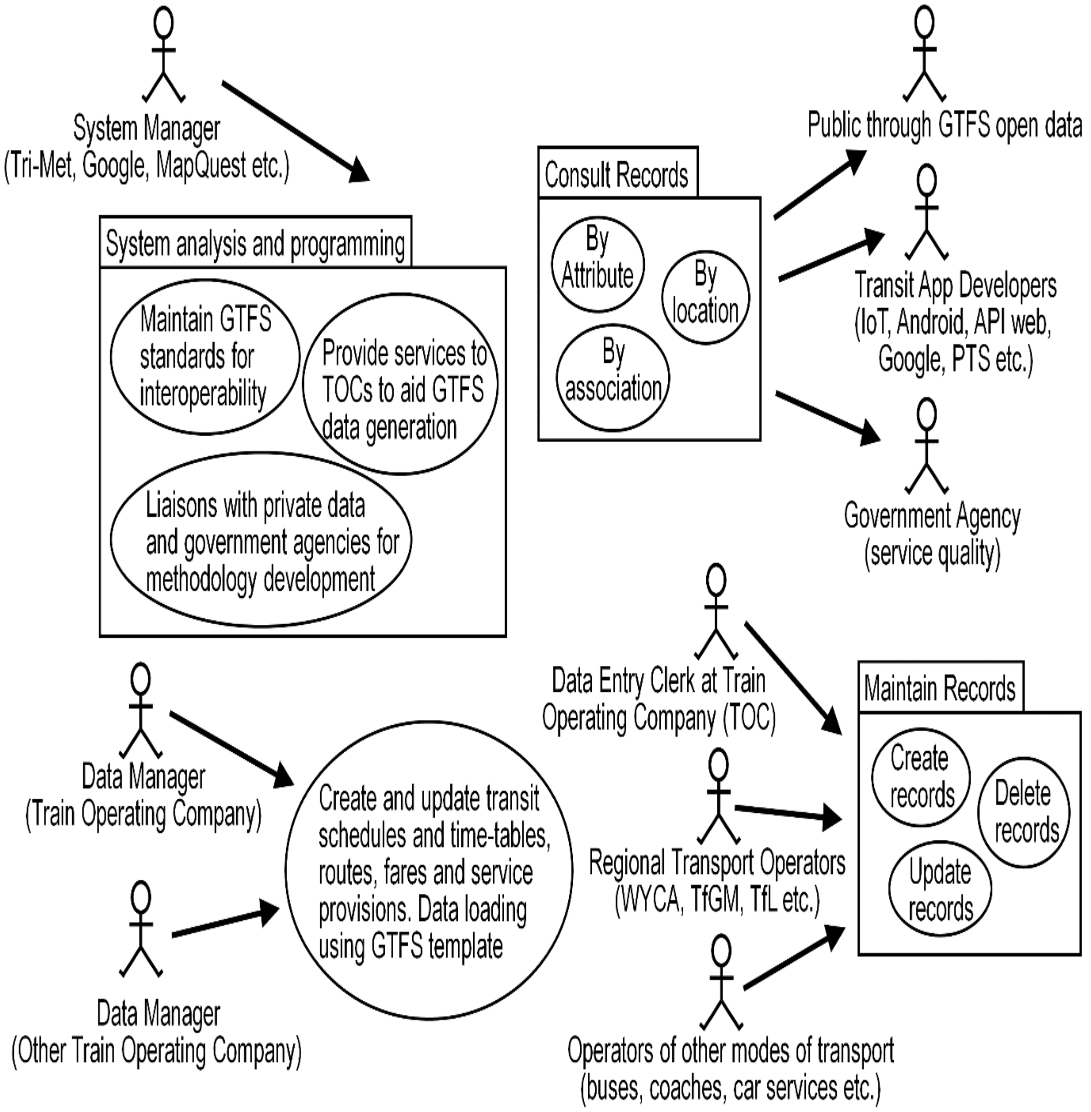

System Manager
(Tri-Met, Google, MapQuest etc.)
System analysis and programming
Maintain GTFS standards for interoperability
Provide services to TOCs to aid GTFS data generation
Liaisons with private data and government agencies for methodology development
Consult Records
By Attribute
By location
By association
Public through GTFS open data
Transit App Developers
(IoT, Android, API web, Google, PTS etc.)
Government Agency
(service quality)
Data Entry Clerk at Train Operating Company (TOC)
Maintain Records
Create records
Delete records
Update records
Regional Transport Operators
(WYCA, TfGM, TfL etc.)
Operators of other modes of transport
(buses, coaches, car services etc.)
Data Manager
(Train Operating Company)
Data Manager
(Other Train Operating Company)
Create and update transit schedules and time-tables, routes, fares and service provisions. Data loading using GTFS template

Figure 6.1 Use case diagram for GTFS information.

The processing procedure for the GTFS[55] information prior to enabling a GIS is summarised in the flow chart of Figure 6.2. GTFS information is regularly published by the Department for Transport (DfT) and the Association Train Operating Companied (ATOC), and the different regional integrated transport authorities (ITA's) facilitate the publication of detailed GTFS information to encourage the development of applications for journey planning, research on accessibility of transport, and for comparing transport service levels.

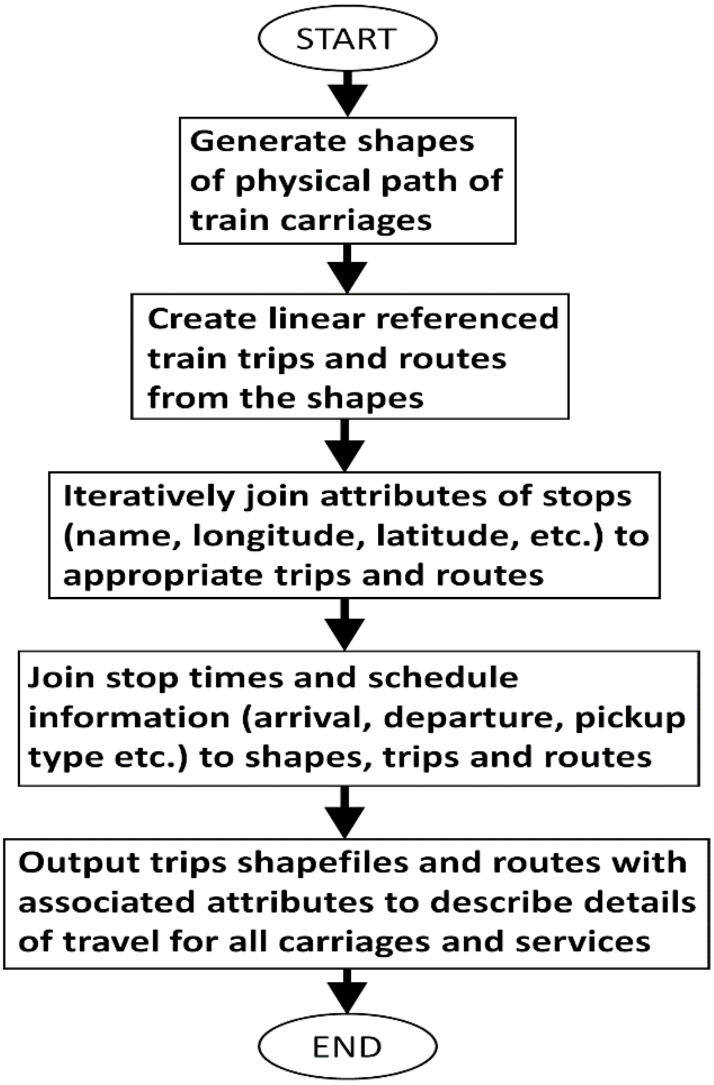

START
Generate shapes of physical path of train carriages
Create linear referenced train trips and routes from the shapes
Iteratively join attributes of stops (name, longitude, latitude, etc.) to appropriate trips and routes
Join stop times and schedule information (arrival, departure, pickup type etc.) to shapes, trips and routes
Output trips shapefiles and routes with associated attributes to describe details of travel for all carriages and services
END

Figure 6.2 Processing steps for the GTFS to embed within a GIS.

[55] Historically the GTFS information was originally conceived at the companies of Google and TriMeT, as a platform for transport agencies to incorporate standard transit data into mapping services and Google Maps. Prior to this, there had not been any standard for public transit data. These standards were originally released as Google, but reference to Google was dropped to attract wider acceptance of the standard.

6.2.2 GIS network dataset

The GIS is the estimable physical model of the transportation network. The connectivity between facets of the transport network[56] are defined within the GIS. For instance, the GIS defines the connectivity between the household Postcode Units, entrance to street, road network, station entrance and transit lines. The connectivity ensures that for instance it is possible to traverse from the transit lines to the road network only via the train station. Further specification of end point connectivity ensures that passengers can travel from one transport link to another only via the link vertex. That way for instance, even if network linkages visually intersect, there would be no possibility of inadvertently facilitating a turn through the intersection. ArcGIS (ESRI, 2016) models are built to form the methodology for creating linear referenced features, to geospatially integrate these features, and to form a consistent dataset. The minor roads are split at points were the Postcode Units and rail stations connect to the road network. The detailed specification of the connectivity adopted is shown in Appendix C.1.

The Connectivity groups created within the GIS make it possible to traverse features within a group, but a linkage to features in another group is only possible via special points and linkages. For example, it would be reasonable to expect to traverse between the minor roads, A- and B-roads, however, connecting onto a motorway would normally occur at a major slip road, junction, or roundabout. As such two connectivity groups are created for the road networks, one made up of the motorway, and another made up of the A-, B- and minor roads. Traversal between these groups would only be possible at appropriate slip roads, junctions, or roundabouts. Similarly it would reasonably be expected that passengers can only access or egress the train transit lines via the train station and station entrance. As such the transit lines and general network of roads would form two separate groups, such that it is possible to traverse between the groups though the link station-entrance and point. Connectivity between groups of features is only feasible though points on the network (e.g. roundabouts, station points, street-Postcode points, junctions etc.), and when these points are associated

[56] The shapefile data on the UK road network are available from Department for Transport (DfT). The shapefile data on the UK railway network are available from Network Rail. These data can also be procured from the Ordnance Survey EDINBURGHDATASHARE 2016. Research Data Service, University of Edinburgh. Edinburgh Data Share., and are also regularly published and updated at www.data.gov.uk.

with a link, they tend to form a separate connectivity group that facilitates linkage and passenger movement between groups. Such methodological homogenous collection of features with rich topological relationships across features is a particular strength of GIS, and this facilitates an ability to scale across any expanse ideally suited for identifying global and local trends.

6.2.3 Combined GIS-GTFS network

ESRI (ESRI, 2016) have developed a GIS tool for adding GTFS to a network dataset (Morang, 2016), by creating a GTFS SQL (Structured Query Language) database that is queried by the GIS in the process of solving each passengers route on the network. The stops and transit lines created from the GTFS are reconciled with the station and track shapefiles procured from the Ordnance Survey (OS) (EdinburghDataShare, 2016). The summary of the steps in creating the GIS network datasets, the GTFS information, integrating these, and solving each passengers route are shown in the flow chart of Figure 6.3. The procedure consists of creating a linear referenced GIS network dataset. Linear referencing enables the location of features to be specified and measured relative to reference locations along the transit linkages. Connectivity is specified between the transport network features, thereby building links across the road network features. The GTFS SQL database is integrated into the GIS network using the ESRI GTFS toolkit (Morang, 2016). Routes are simulated between each OD pair for each passenger in the micro-population, based on time of arrival at station, and time of first train, access distance and mode, etc. (inherited from the NRTS during spatial microsimulation). These attributes inherently encompass the utility maximizing aptitude of each passenger.

The routes for each individual passenger are solved by finding the quickest route traversed, constrained by the transit schedules and impedances derived from the GTFS, and subject to any additional restrictions derived from the LENNON ticketing information. The solved route events for all the passengers are overlaid, merged at intersections, or queried to identify the desired endogenous passenger attributes (like waiting times, time in each train, number of stops, average speed of journey. etc.). The ArcGIS models that have been developed to create the GIS-GTFS network are described in Appendix C. These models are included in the CD accompanying the thesis, and are used to create the network linkages between Postcode Unit, minor roads, A and B roads, motorways, railway stations and rail lines. These models solve the detailed mobility of each passenger on the rail network, to identify waiting times, number of stops, volumes of passengers etc.

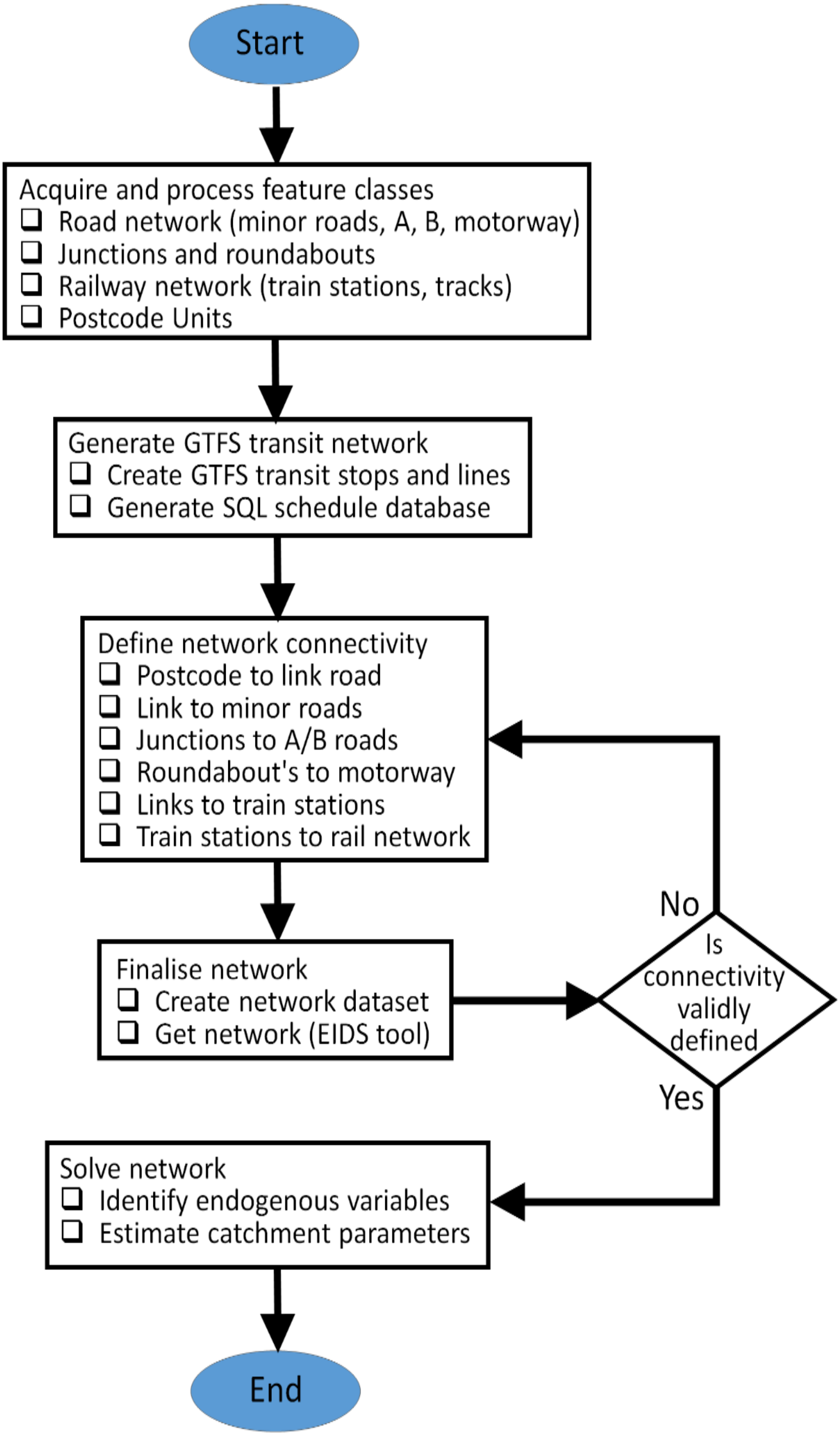

Start
Acquire and process feature classes
❑ Road network (minor roads, A, B, motorway)
❑ Junctions and roundabouts
❑ Railway network (train stations, tracks)
❑ Postcode Units
Generate GTFS transit network
❑ Create GTFS transit stops and lines
❑ Generate SQL schedule database
Define network connectivity
❑ Postcode to link road
❑ Link to minor roads
❑ Junctions to A/B roads
❑ Roundabout's to motorway
❑ Links to train stations
❑ Train stations to rail network
Finalise network
❑ Create network dataset
❑ Get network (EIDS tool)
No
Is connectivity validly defined
Yes
Solve network
❑ Identify endogenous variables
❑ Estimate catchment parameters
End

Figure 6.3 Steps to constrain a GIS network dataset with GTFS information.

6.3 Results from query of GIS-GTFS network dataset

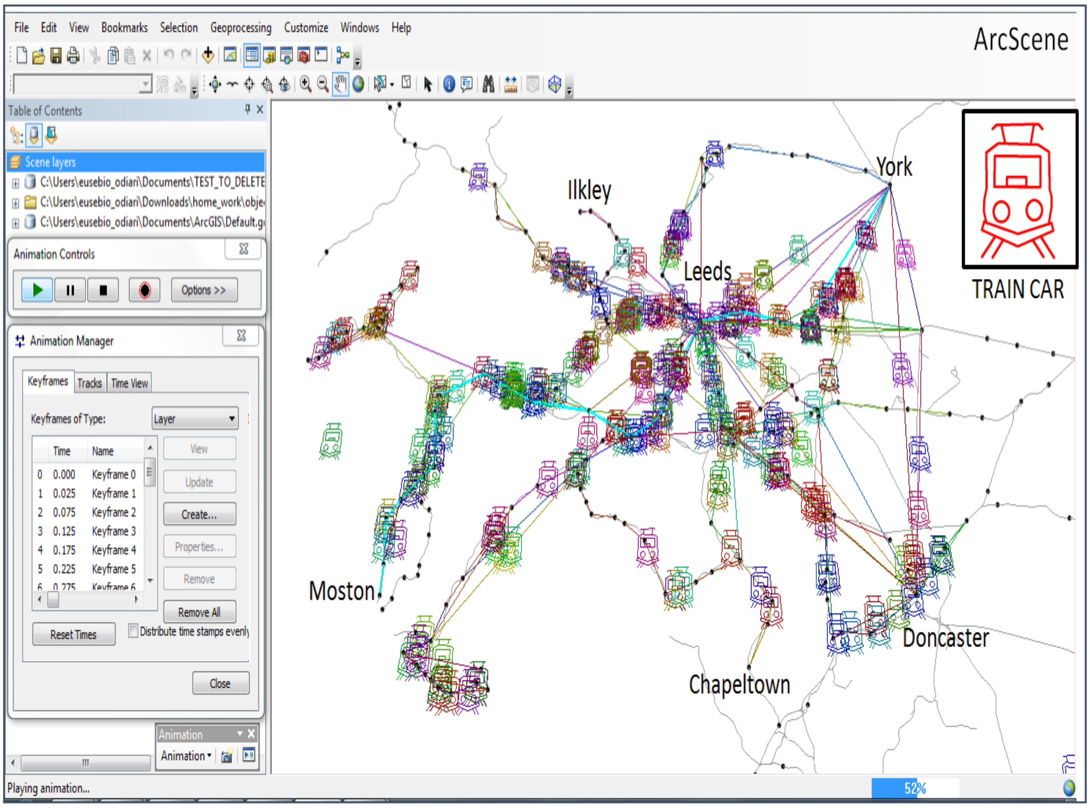
Once the GIS-GTFS network simulation model is developed and built, the prevailing paradigm of interaction became to query the geographic database, and to visualize or animate the results. For the GIS-GTFS network model, the analysis cuts across local and global scales, enabling query and visualisation of micro-level mobility, disaggregate network flows, and aggregate flow volumes. Some of these results are presented in the following sections. Each passenger within the micro-population created from spatial microsimulation has a Postcode Sector of usual residence and final travel destination, as well as time of arrival at station and first train. When the synthetic population are used as input to the GIS-GTFS network, information about the time of first train[57] is used to simulate the passenger's exact flow.

6.3.1 Micro-level mobility on the railways

Figure 6.4 is a screenshot of the trains in mobility within West Yorkshire County (with a 15km buffer applied to include all the relevant stations in the map). These results are derived from the GIS-GTFS network model. The GTFS information analysed are for a period (set by the train operators to be 4 weeks) starting on 7th March 2015 to coincide with LENNON ticketing information available to the project. All the trains on display in Figure 6.4 have unique daily and weekly schedules. The train routes (trips) and trains are time activated, implying that the animation can be synchronised to 24hour daily times. There are 256 unique trips across West Yorkshire, ranging from 1.2km trips (between Swinton and Mexborough) to 108km trips (between Moston and York). To create the service provided by the train operators, the 256 unique trips are replicated at different times within the day and week, to form about 17,350 different train trips. The black dots represent train stations[58] (see Figure 6.4)

[57] To achieve a better and more realistic mobility animation, each passenger can be assigned to their Postcode Unit derived from the spatial microsimulation. The assignment could further be optimized using train station access mode and distance. This however, has not been deemed necessary in this research which focuses mainly on the train mode of transport and within the rail network.

[58] The dots that are not traversed are either railway stations that have been closed or now form part of the transport for Greater Manchester (TfGM) Metrolink service. TfGM flows emanate from outside West Yorkshire and are as such not included in the analysis.



Keyframes of Type: Layer

	Time	Name
0	0.000	Keyframe 0
1	0.025	Keyframe 1
2	0.075	Keyframe 2
3	0.125	Keyframe 3
4	0.175	Keyframe 4
5	0.225	Keyframe 5
6	0.275	Keyframe 6

☐ Distribute time stamps evenly

Figure 6.4 Screenshot of animated trains in West Yorkshire.

Figure 6.5 is a zoomed in screen shot of animated trains (with embedded passengers) on the rail network. On the left in the main window are passenger waiting to board a train. The zoomed in section in the lower-left window shows the train Y53016-IB09 (indicated by the red arrow in the main window with the associated passengers embedded). This is the Blackpool-to-York, travelling from Accrington to York on the 07:02-09:04am train (Route No. 2664, Trip ID 79293, Service ID 451, and Shape ID 979). By querying the GIS-GTFS network model created, such detailed attributes associated with each passengers mobility on the rail network is extracted and included as an endogenous (within network) variable for subsequent use in Bayesian spatial analysis (see Chapter 7 and Chapter 8). Further micro-level results could be the real-time volumes of passengers within a train over the length of trips. Figure 6.6 shows the volume of passengers within a train over the length of trips. Information on number of trains laid out for a service was not available; as such passenger volumes have been reported. Crowding will occur if sufficient trains are not put in service following insight into the volume of passengers in the train. In Figure 6.6, the morning trips (09:07am) are seen to constitute large volumes. Guiseley is seen to have consistent sizeable volumes, perhaps indicative that it is a spur route to hub stations like Leeds.

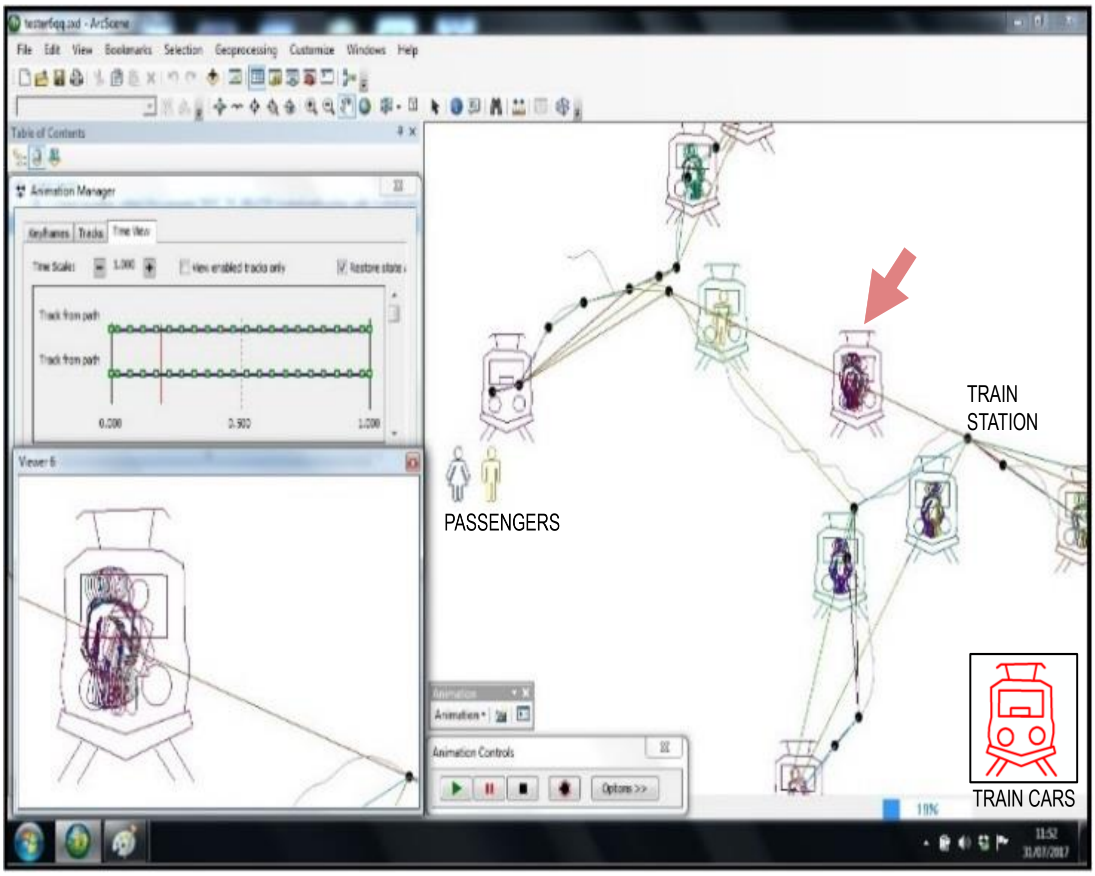


Figure 6.5 Screenshot of animated urban mobility along transit lines.

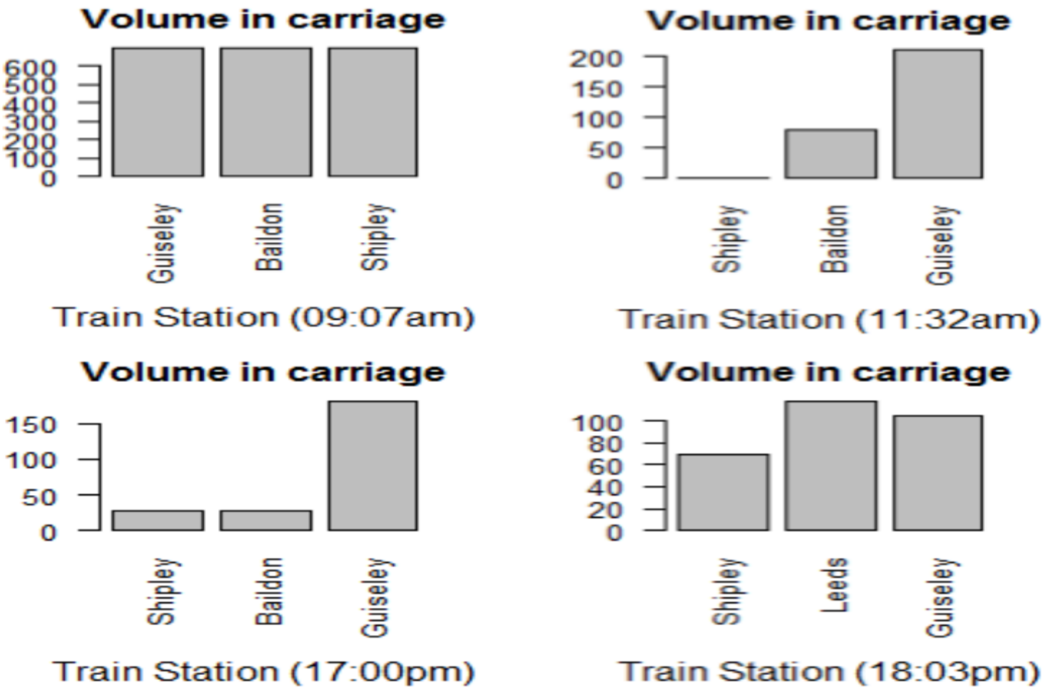


Figure 6.6 Train volumes for different trips during the day.

6.3.2 Disaggregate flow on transit lines

Figure 6.7 shows typical disaggregate passenger flows on transit lines on the rail network. Both the flows depicted on the left and the right of Figure 6.7 are for passengers similarly travelling between Huddersfield and Leeds. The origin and destination of the two flows are the same, but the GIS-GTFS enables particular details of the flows to be disaggregated. The first passenger travelled between 08:00-11:59am (left plot) while the second travelled between 15:00-18:59pm (right plot). The details indicate that the 8:00 am trip stopped at Dewsbury whilst the 15:00pm trip was a non-stop train. Analysis revealed that in the day times the left route enjoys a higher frequency of trains and lower consecutive train arrivals at the stops[59].

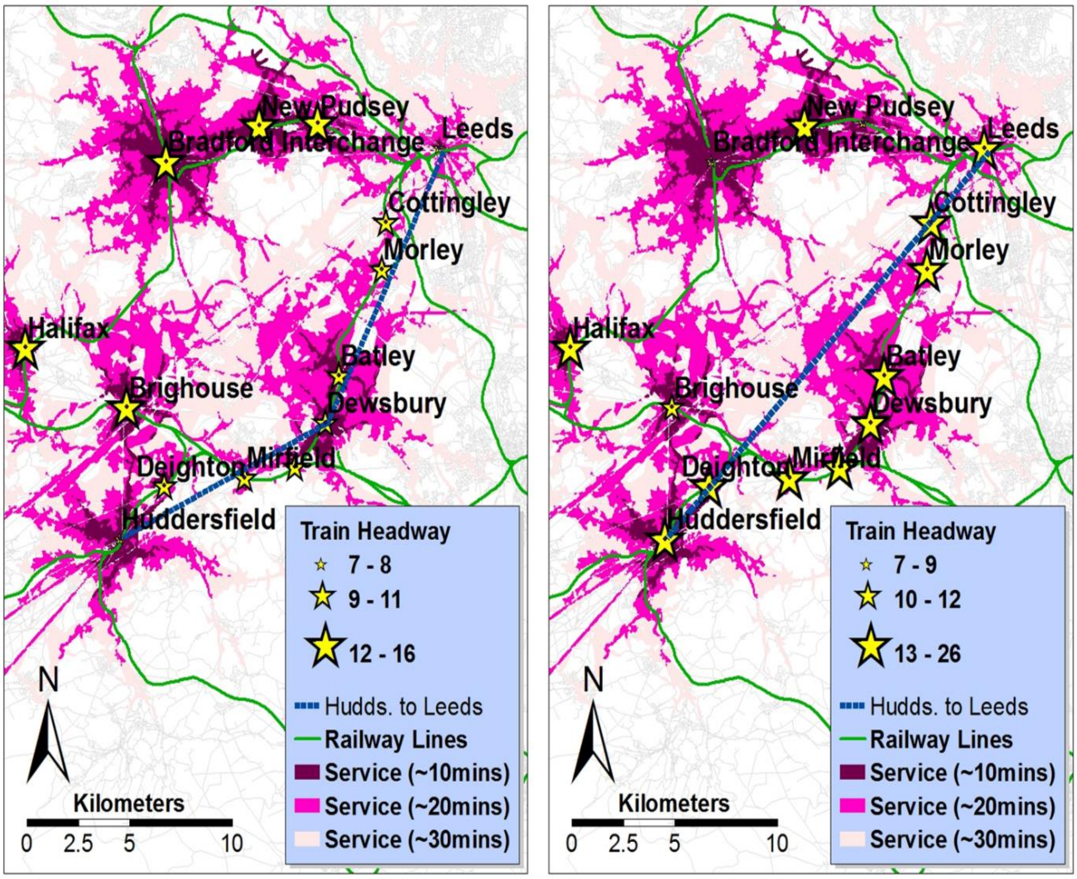


Figure 6.7 Headway and service provision - Huddersfield to Leeds.

[59] The service areas are the same for both periods as mobility outside the network is programmed to be time independent. A GIS-GTFS network dataset can similarly be created for alternative mobility modes outside the rail network, thereby incorporating the characteristics of alternative modes in the analysis. This was considered outside the purview of the immediate project.

Figure 6.8 is another example where a query of the GIS-GTFS network model enabled the disaggregation of seemingly similar mobility on the rail network. In Figure 6.8, the transit journey from York to Skipton can be disaggregated to distinguish the flows. The 11am train (brown line) stops at Shipley, whilst the 8am train (green line) travels direct between Leeds and Keighley, and it is seen that only the 4pm train stops at Cononley. Using the GIS-GTFS model to query the journey reveals distinctive transit routes which are disaggregated by the actual stops that the train makes. Figure 6.9 is further examples of results of a query of the GIS-GTFS network revealing disaggregate mobility. As seen in Figure 6.9 the details are revealed of the route traversed by a passenger travelling between Wakefield Westgate and Leeds. The particular example was chosen to reflect the impact of intermediate stations stated by a few passengers and enumerated in the NRTS. Such convoluted journeys occur as outliers, and perhaps are passengers travelling in groups whereby members of the group are accompanied to particular stations. The GIS-GTFS network model enables the identification of these outlier behaviour for further investigation of the details of travel in a bid to improve the service provisions.

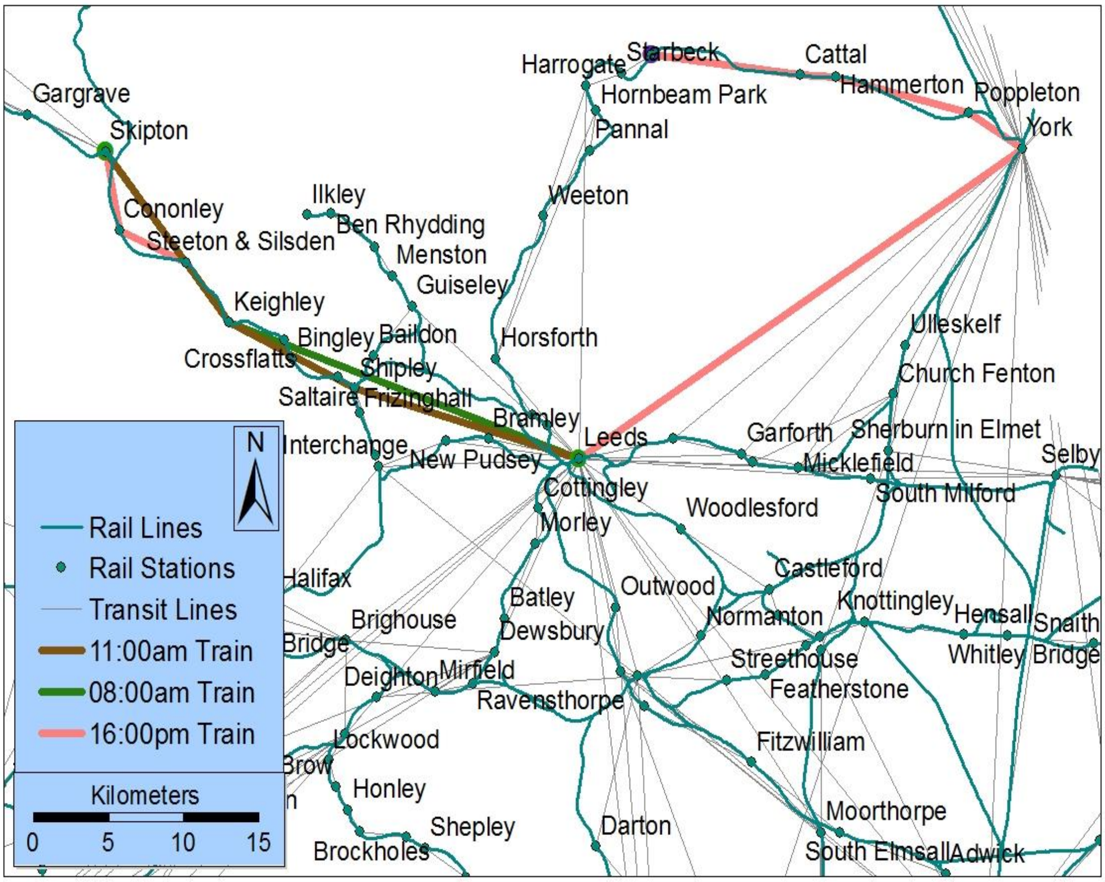


Figure 6.8 Headway and service provision - Huddersfield to Leeds

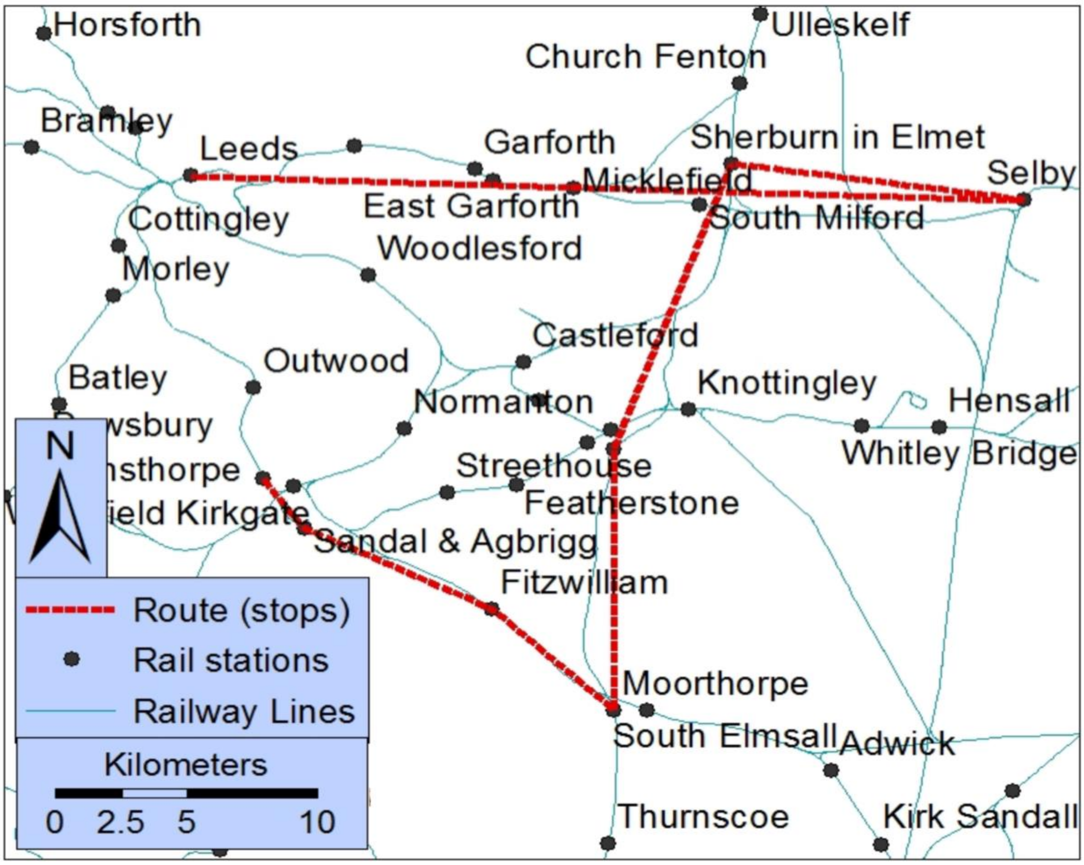


Figure 6.9 Passenger trace - Wakefield Westgate to Leeds station.

6.3.3 Aggregate flow on transit lines

The events associated with the micro-level and disaggregate flows on the network can be overlaid to create an aggregation of mobility. Such information is relevant to create visual summaries of mobility phenomena on the rail network. Typical results would be the detailed volume of passengers within train stations over the period of the day[60] as shown for Leeds train station in Figure 6.10. In Figure 6.10 the upper plot represents the flows aggregated every 1 hour, while flows in 10 minute intervals are shown in the lower-plot. Such analysis enable the passenger flows to be disaggregated right down to real-time mobility. Figure 6.11 illustrates further aggregate passenger flow volumes on the rail transit lines. Overlaying each passenger's network journey in time and space reveals the rich tapestry of mobility on the railways. This detailed space time picture of passenger movement on the railways enables for example the association of the socio-

[60] The volume depicted was derived from LENNON ticket sales aggregated into monthly periods. As such the daily patterns are derived from the volume for the whole month

demographic attributes of passengers to their distinct behaviour on the rail network, facilitating the construction of richer pictures of mobility. A typical overlay of events is shown at close range in Figure 6.11.

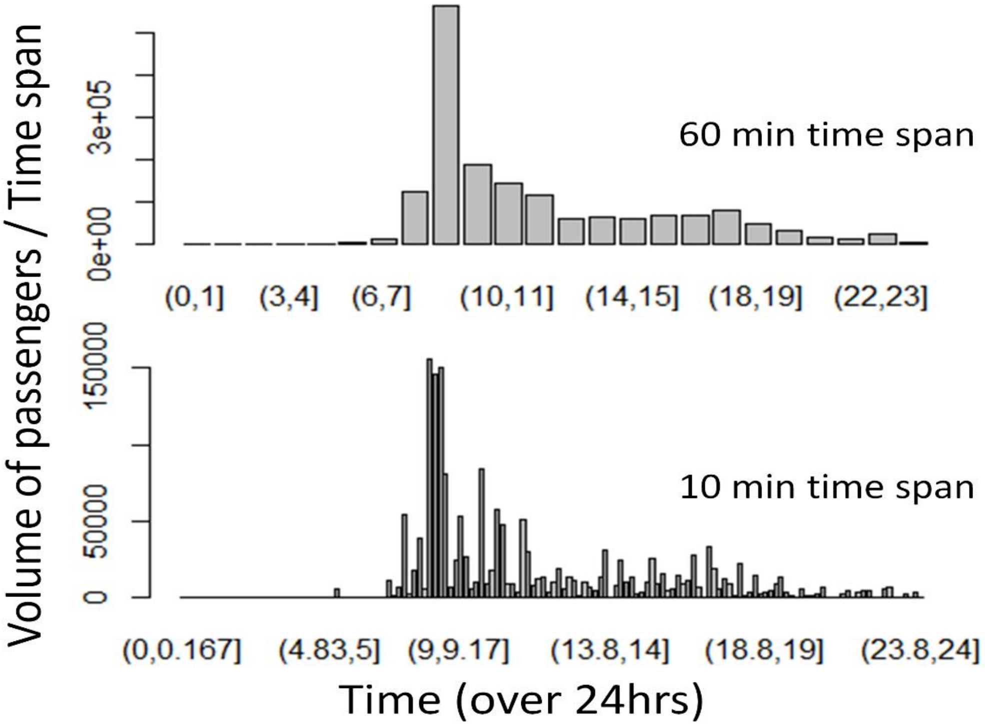


Figure 6.10 Real-time volumes at Leeds Station (P2015/13). Upper plot are flows aggregated every 1 hour, while lower plot are flows aggregated every 10 mins.

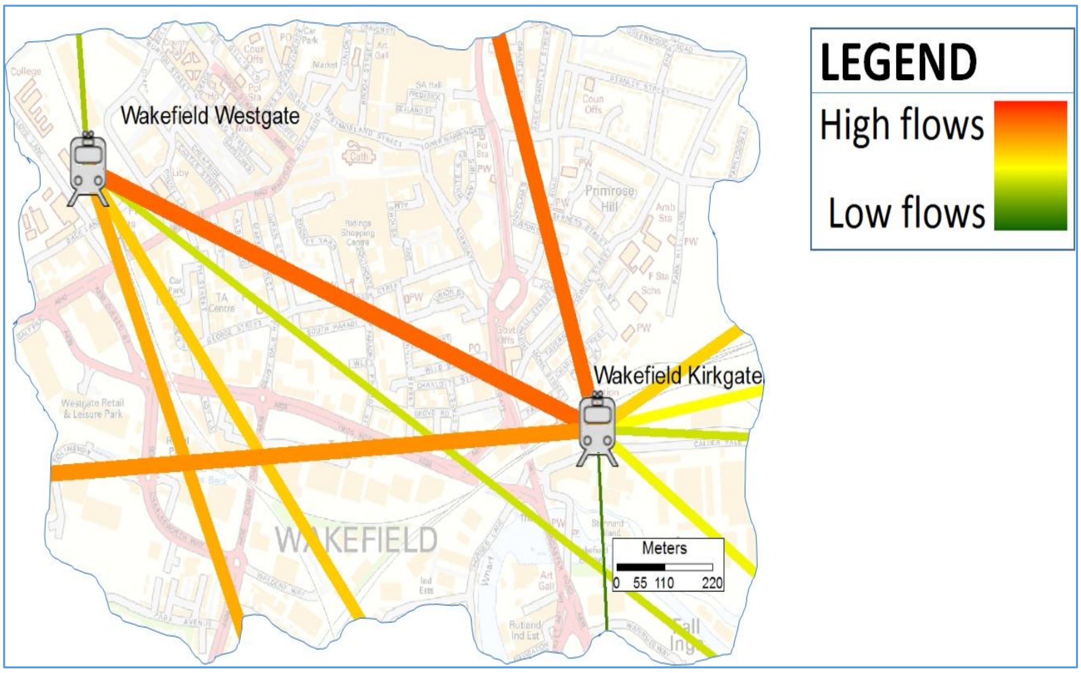


Figure 6.11 Flow volumes - Wakefield group stations.

Figure 6.11 reveals flows in the vicinity of the Wakefield group of stations (i.e. Westgate and Kirkgate). Such local aggregation creates a visual picture of volumes of passengers associated with each transit service that traverses Wakefield. The result shown in Figure 6.11 indicates that a large proportion of the passenger volume is associated with flows occurring between Westgate and Kirkgate. The ticketing data would not always distinguish these flows as the stations are grouped as the same on many ticket products. This research enables an identification of specific flows between Westgate and Kirkgate, as well as disaggregating these flows temporally if desired. The indication from these results is that a disruption between Westgate and Kirkgate will have most effect on the volumes of passengers in the vicinity. Figure 6.11 is a local view of the aggregate mobility volumes along transit lines, whilst Figure 6.12 shows a global view of passenger volumes along transit lines in the West Yorkshire study area. Figure 6.12 for instance can serve as a visual summary of service fulfilled by each transit line, as such information is not readily available from LENNON. The LENNON ticketing data does not include information on time of passenger travel; however, the GIS-GTFS model enables a deduction of aggregate or disaggregated temporal flows of each passenger at different locations on the network.

A further paradigm for display of passenger volumes is shown in Figure 6.13. In this instance the aggregate global view of passenger volumes associated with each geographic railway line are displayed (as opposed to the straight transit lines between stations). In the West Yorkshire study area, Figure 6.13 reveals that aggregate flow volumes tend to radiate outwards, decreasing from Leeds Station. This reveals that Leeds Station is the hub of rail mobility in the county, and service provision would need to have core strengths in the vicinity of Leeds. Such information can be deduced form the ticketing data, but the strengths of the GIS-GTFS model lies in its ability to enable visualisation of such query results on mobility on the rail network. A temporal visualisation of aggregate flows throughout the day would reveal a map of passenger flow volumes at different stations and locations throughout the day. Such temporal information would not be readily available from the LENNON ticketing data. Results shown in Figure 6.13 can further be used to good effect to reveal volumes of transit trains required to ply different segments of the rail lines. Such information can further be useful in estimating the load that sections of the rail lines might have endured, as part of a maintenance scheduling tool.

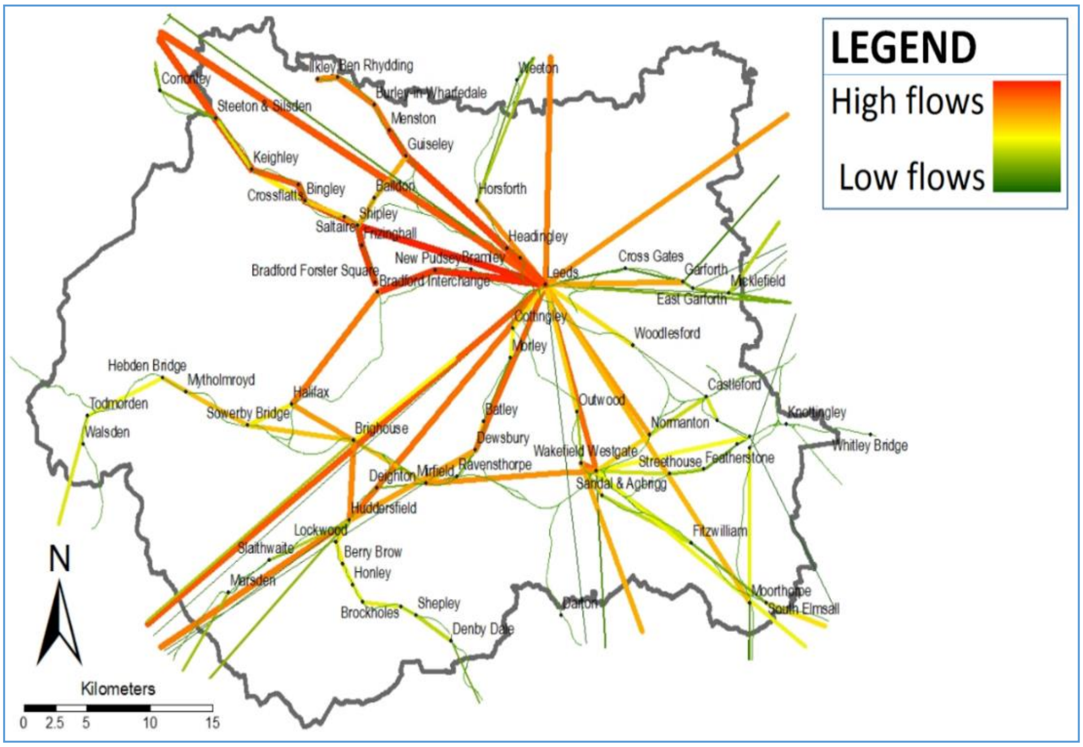


Figure 6.12 Flow volumes- transit lines in West Yorkshire.

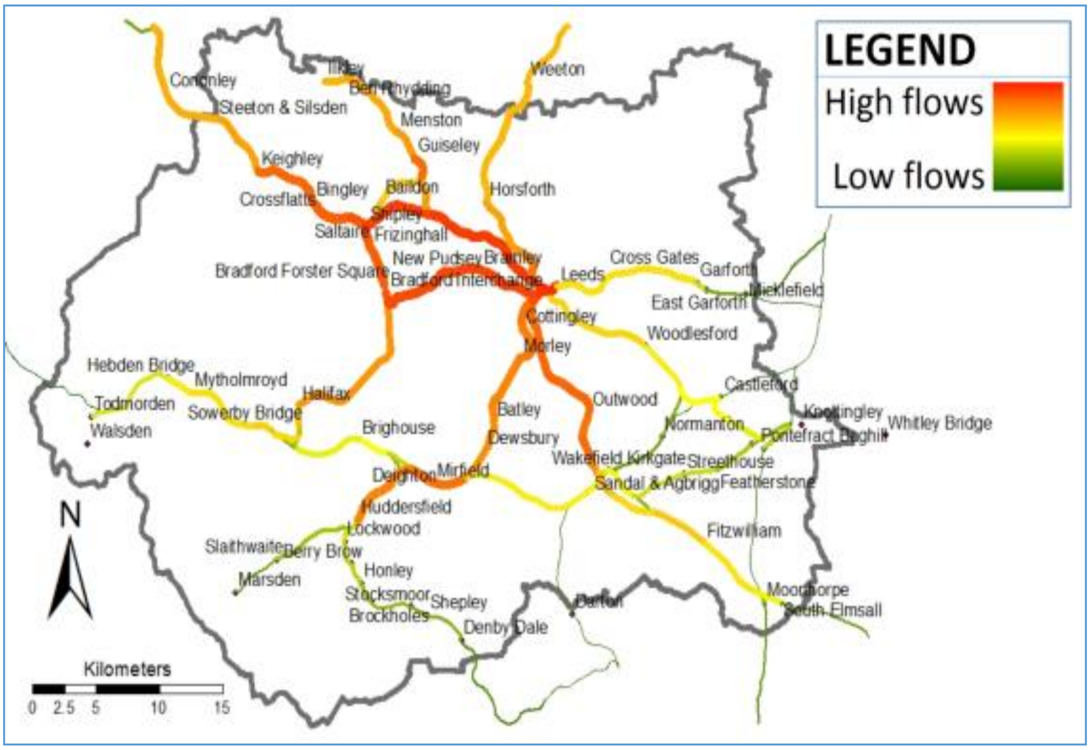


Figure 6.13 Flow volumes – rail lines in West Yorkshire.

A further aggregate analysis yields a measure of regional access to rail services shown in Figure 6.14. The image reveals a typical in-house transport coverage index (TCI) which is a combined measure of volume of trains at a station, density of rail stations in a defined road-network buffer, maximum waiting times at each station for train service and area of non-overlapping service coverage. The equation developed for the TCI is shown in Equation 6. These measures of transport coverage are predefined network performance attributes, as the network operators pre-define train routes and the associated number of stops, waiting times etc.

$$TCI = \frac{(No.\, of\ Trips)(Stops\ in\ Area)}{(Max.\, Waiting\ Time)(Overlap\ Service\ Coverage)} \qquad (6)$$

It would be desirable to have TCI values uniformly distributed with wide coverage across the geography (West Yorkshire County). The TCI is useful as it is indicative of areas where service management requires attention due to heightened mobility activity. The TCI is also useful as an indicator of areas with limited access to rail services (the blank white areas with zero TCI). The GIS-GTFS network model automatically deduces TCI values for the geography, and particular TCI values can be associated with each passenger as an additional attribute for use in mobility analysis.

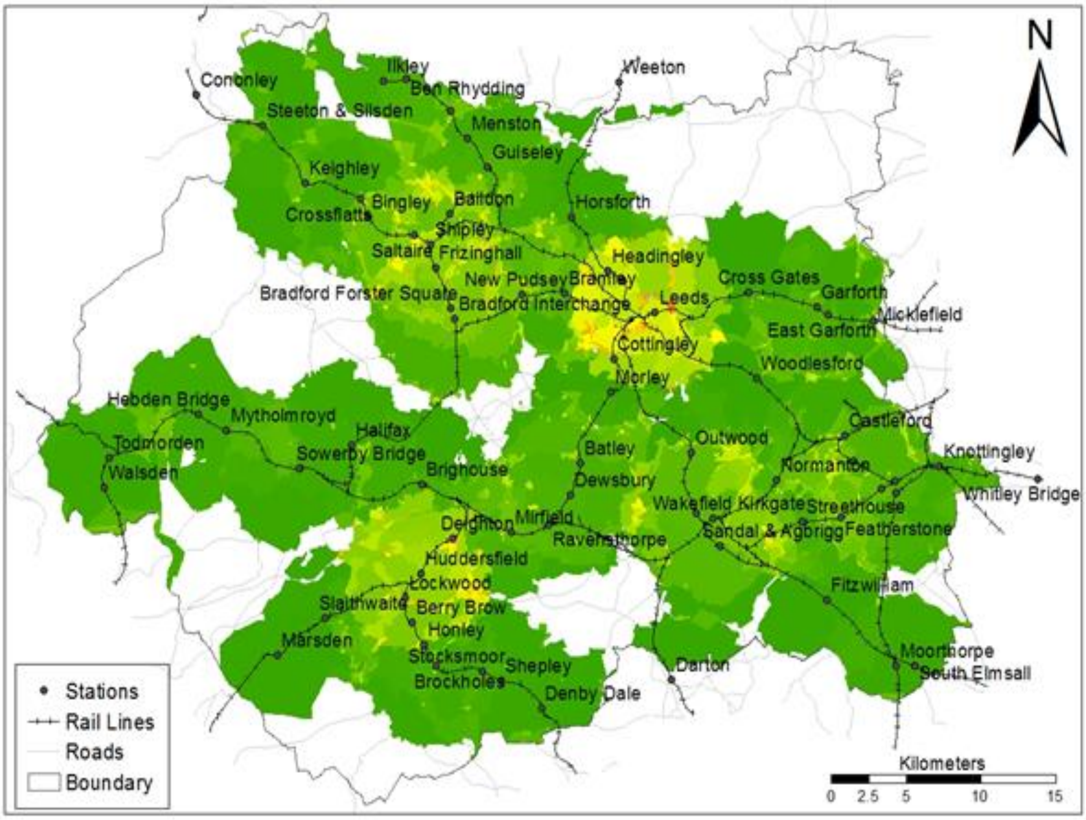


Figure 6.14 In-house TCI (Index) in West Yorkshire.

6.4 Discussion of GIS-GTFS network simulation

In this chapter, the micro-mobility of passengers on a UK railways system is replicated by creating a GIS-GTFS network. This enabled the reconciliation of a simulated spatial micro-population outside the network with a within-network micro-mobility population. This combination produces a representative micro-level population. The GIS network model incorporating detailed transit schedules is used to identify the context of ticket use, revealing the additional endogenous passenger attributes such as waiting times the passenger might have endured, number of stops, average transit speeds, volume of associated passengers (a proxy for crowding), etc. Such individual-level revealed information augments current industry estimates of passenger attributes and activity on the railways. These were hitherto based on traditional low sample ratio stated preference surveys that were prone to survey biases. The research as such has significant impact on the quality of inputs available to strategic, tactical and operational rail industry planning models. The GIS-GTFS model is potentially useful to network operators for management, maintenance and interventions assessment on the UK railways. The representative micro-level population can aid in creating segmentations on railway customer satisfaction by linking in satisfaction surveys. Such segregation can aid in answering pertinent policy questions related to how concessions might affect different socio-demographic groups, to identify how different population groups are affected by the various service provisions. The GIS-GTFS network model can additionally reveal the volume of passengers and transit train weight that each segment of track bears in service, thereby enabling more holistic track maintenance scheduling.

The GTFS information published by ATOC and DfT (ATOC, 2017) for the jurisdiction of the West Yorkshire Combined Authority does not incorporate fares information, as such the fares variable adopted was that included in the LENNON dataset. The plausibility of further optimising the solved passenger routes based on derived fares was not incorporated, but considered as future research effort. Conventional shapefile and database file formats do not support date and time in the same field, however the ArcGIS geodatabase feature enabled the incorporation of precise times of travel (in seconds) for each passenger, thereby enabling the replication of realistic mobility. The flows analysed in this chapter (and in the thesis) consists of flows originating and terminating in West Yorkshire County. This restriction has been imposed as the LENNON ticket availability was restricted to West Yorkshire County. Similarly also, the NRTS availability was limited to the region of Yorkshire and the Humber. A large volume (in the region of 20~35%) of National Rail flows through West Yorkshire either emanate or terminate outside the County, and this affects the volume of passengers simulated, since the LENNON tickets formed the final target dataset.

Internal validation of the GIS-GTFS was conducted by comparing the flow times deduced for each passengers movement with that deduced from the current Google software results (Cardno and Mulgan, 2003, Rafiah et al., 2004). These results coincided with Google for the entire 5% sample of the individual mobility's solved using the GIS-GTFS model. This exercise was repeated several times over a period of 1 year, and on each occasion the results deduced from the GIS-GTFS model coincided with those from the current Google data. This provided the internal validation of the GIS-GTFS network model. Unfortunately Google does not provide historical passenger travel plans, and as such it was not possible to compare Google with results deduced from the GIS-GTFS model using historical GTFS information. It was deemed sufficient that current GIS-GTFS and Google results coincided for all cases tested over a period of one year.

A scheduled smart card data would have been ideal for external validation of the GIS-GTFS model; however, such data was not made available in the time frame of the research project. The availability of smart card data in the future would provide an opportunity for a more comprehensive external validation of the GIS-GTFS model results. The GIS-GTFS network simulation developed in this research is reliant on a representative simulated population (in this case from the m-IPF, but also could have been from SA). The strategy however, does not assume a model for assigning passengers to the network traffic as is the case with traditional multi-agent traffic modelling software (Anderstig and Mattsson, 1998, Gao et al., 2010, Horni et al., 2016), instead the richness of the endogenous and exogenous attributes in the synthetic population is exploited. The synthetic population contains information on passenger arrival time at train station, first train, intermediate stops, final destination, ticket restriction, access and egress mode and distances, as well a plethora of addition information to accurately assign the passenger to the network traffic. Any resulting space-time congestion simply reflects the actual situation on the network, and precludes the need for the complexities of optimizing the utility of passenger mobility in traditional multi-agent modelling. Such micro-level demand and supply information are potentially useful to network operators for planning, management, maintenance, and interventions assessment on the railways.

Chapter 7
IMPUTATION OF BEHAVIOURAL DATA

Having used the m-IPF (spatial microsimulation) and the GIS-GTFS network model to enhance the LENNON ticketing data (in Chapter 5 and Chapter 6 respectively), an individual-level attribute-rich dataset is yielded. The attributes from m-IPF include those variables belonging to individuals surveyed in the NRTS, as well as product information on individual tickets within the LENNON database. These attributes are optimised by reconciling with Census information on aggregate passenger mobility interaction. The GIS-GTFS network model on the other hand further enhances the individual-level data from the m-IPF by using the simulated passengers as input into the logistical GIS-GTFS rail network, constrained by the transit schedules. Such a logistical network enables the identification of the context of each passenger's mobility on the rail network. These additional contextual attributes like waiting times the passenger might have endured, number of stops, crowding in train, length of journey, etc., all add to enhance the range of individual-level attributes associated with each passenger. The resulting enhanced LENNON ticketing data is ideally suited to identify drivers of mobility on the railways, aimed at predicting individual movement patterns, considering frameworks relating population mobility to demographic characteristics, relative location, time of the day, and other likely drivers of behaviour and interaction.

Despite the rich set of exogenous and endogenous attributes associated with simulated passengers from the m-IPF and the GIS-GTFS model, certain passenger behavioural attributes are still missing, like daily rates of season ticket use (so called journey factors), length of use of flexible period tickets, and precise passenger flows to group stations[61]. These further missing values need to be imputed to create a comprehensive estimate of passenger demand.

[61] It is noteworthy to point out that the precise flows associated with regional multi-modal Passenger Transport Executives (PTE) tickets are unknown (and hence missing), but these are conceptually similar to Group Station flows (which are inherently missing). The similarity is that both PTE and Group Station ticket types restrict passenger flows to a pre-defined number of known stations. PTE tickets were however not available to the project, and as such have not been included in the analysis.

7.1 Introductory background

In order to bring out the imputation aspect more strongly, this section highlights the specific deficiencies in the LENNON ticketing data which this research seeks to address. These deficiencies are missing values in the variables for journey factors, ticket validity and grouped stations, and these three problems will be addressed in this chapter. In this section (Section 7.1), the background to each of these items is given, indicating by way of simple examples how they arise, and why they are important. Further, examples are used to explain the concepts behind the methodologies developed to resolve the missing values by way of developing imputation strategies.

7.1.1 Missing values

If for example it was established from the m-IPF and GIS-GTFS that a passenger associated with a particular ticket in the LENNON database was also surveyed during the NRTS, and was found to travel regularly with a 7 day season ticket, between a specific origin and destination station. The m-IPF and GIS-GTFS strategies enabled each LENNON ticket to be assigned to a passenger, however, there is no indication of whether that passenger who might have been assigned a season ticket, used the ticket many times in the day. Such a passenger for instance could have used the season ticket thrice in the day, but the m-IPF and GIS-GTFS would not identify this information. Apart from the daily rate of ticket use, the 7 day season might have been used only for 5 days (for a 5 day working week), in which case the 7 day ticket would have been used for only 5 days of a 7 day validity tenure. Such information would be necessary in building a compound picture of passenger demand. Apart from fixed period season tickets (like the 7 day season ticket or monthly season ticket), there are flexible period tickets which could for instance be valid for between 30-60days, without a firm indication of the number of actual days the ticket might have been used despite the 60 day maximum limit. Such information would be necessary in building a richer picture of passenger mobility.

Apart from daily rates of ticket use and validity periods of flexible tickets, the precise passenger flows to those stations clustered into groups (so called group station flows) are missing. In the UK railways network, tickets are typically issued to individual origin and destination stations. In some cases however, there are a number of stations in the same locality, whereby a passenger may desire to travel to one of the stations and return from another station in the same locality. To accommodate this particular passenger service demand, British Rail (BR) introduced a series of station groups. In West Yorkshire, typical examples of group stations are Bradford BR (made up of Bradford Interchange and Bradford Forster Square, see Figure 7.2) and Wakefield BR (made up of Wakefield Kirkgate and Wakefield Westgate). Tickets are for example issued for destination station Wakefield BR, and the passenger may choose to alight at Kirkgate and return from Westgate. It would be desirous for the train operators to for instance decide whether to lay out more transit trains for journeys from Bradford Interchange or for those from Forster square, and also to decide the temporal nature of the demand across these stations within the same group in order to achieve an optimal service provision closely tailored to the demand. In essence, for the purposes of mobility analysis, a richer picture of mobility entails establishing the precise flow volumes to each of the stations within a group.

7.1.2 Concepts adopted

As intimated earlier, there are several methods developed for imputing missing data values in datasets (Allison, 2001, Efron, 1994, Little and Rubin, 2002, Rubin, 1976). These methods range from simple ad hoc or logical-rules-based infilling methods to the more advanced principled imputation methods. The ad-hoc or logical-rules methods are typically based on estimating missing values by interpolating between existing observed values. These simple methods developed rules derived from an understanding of the data creation process and scenario. They are called ad-hoc because they are not general solutions, but are instead problem specific solutions created for a particular purpose. The principled imputation methods are based on more sophisticated rules typically involving creating a model for the data generation process or creating an algorithm that enables a heuristic discovery of the data generation process.

This research is focused on developing a principled imputation method that complements current logical-rules based industry infilling methods. The Bayesian imputation strategy is the choice method developed because it is seen as a more robust application, which includes not just a model of the data generation process, but also a model for the mechanism of missing data. This comprehensive imputation approach implies that the missing passenger behaviour can be inferred by replicating the context in which the mobility data was generated. In the Bayesian framework, the mobility model that describes the data generation process is solved simultaneously with the imputation model that describes the statistical mechanism of missing data.

The choice of Bayesian imputation and analysis is three fold: First, the Bayesian strategy is based on and driven by the powerful MCMC[62] technique, which is an established platform (Geyer, 1992, Hastings, 1970, Metropolis and Ulam, 1949, Robert, 2004) for solving complex intractable models (like those associated with urban mobility on the railways). The MCMC is based on inferring posterior parameter distributions from the likelihood values deduced from available data, alongside a prior intuition about the parameter distribution. Secondly, the Bayesian framework provides a method for imputing missing values within the context of a model of the data (mobility) generation process. As such, the imputation and analysis can be conducted in the same step, providing for a robust methodology. Thirdly, there exists tools within the BUGS implementation of the Bayesian strategy (Lunn et al., 2012) that enables more realistic yet parsimonious models to be created, thereby better reflecting a realistic mobility scenario. The intuitions behind these specific three strengths of the Bayesian strategy are hereby explained further in a non-technical manner.

The MCMC is a stochastic algorithm developed to facilitate sampling the posterior distribution of parameters of a model (as intimated earlier), however, irrespective of the model complexities (non-linear, hierarchical, mixtures, etc.). In the Bayesian framework, the parameter values deduced are distributions (instead of fixed values with confidence intervals as is the case in traditional frequentist statistical analysis). The intuition behind a parameter being a distribution is that it makes provision for a range of values for the parameter as reflected in the variability that would be likely based on the data variables being analysed. To illustrate further, imagine a large classroom made up of Chinese students and German students in equal proportion, with age ranges randomly scattered across say 5yrs to 15yrs. If it is desired to deduce a parameter governing the relationship between age and height, the traditional frequentist paradigm would be to regress the data variables (age and height) and infer a statistical parameter value say 1.5 (indicating that on average a unit increment in age yields an increase in height by 1.5 times). In the frequentist approach a 95% confidence interval in the results (say 1.2 to 1.7) is additionally specified. The Bayesian approach is different as the parameter value reported would be a distribution, perhaps

[62] The Monte Carlo Markov chain (MCMC) methods constitute one of the most important algorithms of the 20th Century useful in solving a wide range of intractable optimisation problems typically found in nature.

indicating two peaks at 1.3 and 1.5 better reflecting the different growth rates of Chinese and German students (Duan et al., 2013). The parameter distribution in this case might look like two joined normal distributions with modes at 1.3 and 1.5. The portion of the normal curves with lower heights away from the modes reflects the less likely parameter values based on the scatter in ages and heights. This simple example highlights the intuitive nature of a Bayesian modelling strategy for solving complex problems (in this case a mixture and a hierarchy of students).

The second aspect that makes Bayesian imputation and modelling attractive is as follows: The Bayesian modelling implemented using the BUGS language (Goudie and Mukherjee, 2016, Lunn et al., 2012) is declarative and not procedural like traditional software languages. This means that the order in which components of models or whole models are specified is not relevant. An advantage of the declarative feature is that it is possible to simultaneously specify a mobility regression model (representing the relationship between an outcome variable and pertinent covariates) alongside an imputation model (describing the statistical mechanism of missing data values). The declarative nature of the BUGS modelling language then enables the simultaneous MCMC solution of the mobility model and missing data model. In essence, this results in effectively identifying the missing values within the context of the mobility model which represents the circumstance and conditions through which the data was created in the first place.

The third aspect that makes Bayesian imputation and modelling attractive is as follows: The realistic yet parsimonious modelling ability within the BUGS Bayesian framework is achieved through the availability of simple but effective constructs in the modelling language. For example, in the case of the large class of Chinese and German students, assuming that 15% of student heights and 20% of student ages are missing. Assuming also that our experience of measuring student heights and ages have indicated that the scale used to measure the student height only gets as far as 1.7m, and that pupils less than 6yrs are observed to be less likely to be able to articulate their ages. Our indication then would be that the missing heights are definitely those above the 1.7m limit of the height bar, whilst the missing ages are more likely to be ages less than 6yrs. It is intuitive to expect to better infer the missing values by creating an additional missing data model that informs our Bayesian regression model about the additional information on the missing heights and ages. To implement these we recall from Chapter

6 that the Bayesian framework is based on that a posterior estimate is equal to (or strictly proportional to) the product of our prior intuition about a parameter value and the likelihood contribution of the values within the data points. In the case of censored heights, we can assume the missing height data have a standard (perhaps a normal) distribution and also include a statement that the likelihood contribution of the missing heights are larger than those for heights near the limit of the scale. There are simple inbuilt constructs within the Bayesian framework to automatically implement this censoring construct, and the strengths of the Bayesian BUGS is the availability of many simple but effective constructs to enable a better replication of realistic scenarios. In the case of the missing ages, we may have additional knowledge that the minimum age for entry into school is say 5yrs. In that case the missing ages would in any event be truncated to a lower limit of 5yrs. This additional missing data information can simply be modelled using the truncation construct in BUGS Bayesian modelling, creating a model that better represents the real classroom phenomena.

7.1.3 Summary

In this chapter, having presented the deficiencies in the LENNON ticketing data in simple terms, and explained the concepts and intuition behind the methods developed for imputation in simple terms, a review of data imputation strategies are presented. Emphasis is placed on imputation of data that are typical of those available in the railways industry. Simple ad-hoc logical-rules-based and more advanced principled imputation strategies are reviewed. The principled Bayesian imputation is then developed as a complement to the current logical-rules-based imputation strategy used in the UK railways industry to impute the LENNON ticketing data (SDG, 2014, Taylor, 2013b). A simultaneous Bayesian analysis strategy is developed for application to railways mobility in the West Yorkshire study area. The methodology for imputation and mobility spatial analysis highlights the Bayesian hierarchical modelling strategy, as well as the procedure for applying constraints to parameters and data values in mobility models. These constructs enable more realistic models to be created, as well as facilitate the incorporation of industry experience on estimates of parameters and data values. The Bayesian strategy is shown to have advantages in being straightforward and intuitive in applications, setting the stage for making predictions and inference in a number of case studies. The infilling strategies developed in this chapter facilitate the creation of a requisite dataset for used in spatial mobility analysis in Chapter 8.

7.2 Review of imputation strategies

A number of data infilling (also called imputation) methods have been developed and evolved over the years as widely reported in the literature (Allison, 2001, Graham, 2009, Heckman, 1979, Rubin, 1976, Rubin, 1996, Schafer and Graham, 2002, Stekhoven and Bühlmann, 2012). These methodologies can broadly be classified as ad-hoc methods, heuristic techniques, and parametric (including Bayesian) methods. The name ad-hoc is used in a figurative manner, and simply serves to highlight that the ad-hoc imputation methods are developed for particular problem-specific applications. Ad-hoc methods (Sterne et al., 2009) replace the missing values directly based on an assumption about knowledge of the missing values. Heuristic (Stekhoven and Bühlmann, 2012) and Parametric (Allison, 2001) imputations are principled methods as they do not assume knowledge of the missing values, but use information and the context of the observed data as the basis for estimating the missing values. Heuristic methods develop practical evolutionary learning algorithms or exploit established learning algorithms to estimate the missing data values. Parametric methods assume a prior parametric model and distribution for the data, and these are assumed to explain the mechanism through which the data was created, and use this as a basis for estimating the missing data. Heuristic methods are non-parametric with no inherent assumptions about a model or the distribution of the data. In practise, ad-hoc, heuristic and parametric methods variously employ statistical tools, Bayes theories, stochastic and iterative strategies in the processes of deriving a solution to the chosen methodology.

7.2.1 Ad-hoc imputation

Ad-hoc imputations methods involve strong assumptions which if not met invalidate the inferences that are based on the data (see also Chapter 2, Section 2.7.1). Common ad-hoc methods (Bartlett et al., 2014, Grace-Martin, 2018) are the list-wise deletion, simple means, last observation carried forward (LOCF), inverse distance weighting (IDW) and Kriging, and logical rules based imputation (ORR, 2014c). Variants of these are reported widely in the literature (Allison, 2001). List-wise deletion drops units (i.e. records) in the dataset that have any missing item. The simple means method replaces a missing data item with the mean of the associated variable, while the LOCF method replaces the missing value (item) by the last observed value of that variable (Donders et al., 2006). The IDW and Kriging (and Co-Kriging) are interpolation techniques also used for data imputation. The IDW and Kriging (Kbiob, 1951, Shepard, 1968) are weighted averaging methods, and ad-hoc since the missing point is imputed directly by assuming a relationship to nearby observed points. Current industry infilling practice applied to the LENNON ticketing data has been to use proxy trip rates for season tickets, and ad-hoc logical rules about where multi-modal tickets are likely to be used. The SDG[63] assignment for PTE tickets (SDG, 2011, SDG, 2012, SDG, 2016, Taylor, 2013b) assumes that the cheapest alternative is chosen, assuming that for example MetroCard Zone 1-5+6 (sold by regional operators) are unlikely to be used on a Harrogate-Leeds journey, as a Rail Settlement Plan (RSP) ticket (sold by UK National Rail) would be cheaper and the more likely ticket choice for that journey.

With ad-hoc methods, data are completely deleted or effectively made up. When data is simply deleted the inferences are biased unless the unlikely assumption of complete randomness (MCAR[64]) in the data is satisfied. On the other hand, effectively made up data creates invalid statistics and cannot be recommended as a robust strategy (van der Heijden et al., 2006). As a result ad-hoc methods only produce useful results when the proportion of missing data is small and <5% (Acuna and Rodriguez, 2004) or when knowledge of the mechanism of missing data is firmly established. In the absence of such knowledge, principled imputation strategies would better leverage and maximize the advantages in the accuracy and precision of revealed consumer data (and the so called big-data), in particular when applied to urban mobility and movement patterns analysis as researched in this thesis.

[63] Steer Davies Gleave (SDG) are the current ATOC consultants developing the PTE infill strategies for the rail industry using methodologies introduced by Mott MacDonald to create the Rail in the North Demand and Revenue Model and the annual Origin-Destination Matrix.

[64] In a randomized control survey, only a proportion of the population is surveyed or sampled (as specified by the sample ratio). This sample is statistically representative of the population. Those parts of the population that are not sampled (i.e. unobserved) are as such missing. In a randomized control experiment the missing part (i.e. those that were not surveyed) occur entirely at random and are said to be missing completely at random (MCAR). Note however that there are more technical MCAR definitions that are widely available in the literature, and these require that the missing part are independent of all the variables that define the investigated problem, as well as the unobserved parameters of the problem.

7.2.2 Heuristic imputation

Many consumer datasets pertaining to passenger travel behaviour reveal the idiosyncratic variability in passenger behaviour which often reflect in a broad range of phenomena. Non-parametric heuristic imputation methods are often suited in situations where an exhaustive search for an optimal solution is not practical, but a satisfactory indicative solution would suffice. These are typically in complex problems like those posed by the non-linearity, mixtures, and hierarchical phenomena associated with mobility interactions. Parametric imputation methods depend on constructing models based on intrinsic prior knowledge of systems that are not fully understood with no guarantee on results from such typically cyclic development exercises. The non-parametric strategies on the other hand are iterative in nature, and rely on large datasets making them suited to novel large multivariate consumer datasets. Many non-parametric imputation strategies (support vector machine (SVM), K-nearest neighbour (KNN), etc.) have been developed over the years (Mazumder et al., 2010, Troyanskaya et al., 2001). Random Forest (RF) imputation for example are a non-parametric strategy for mixed data types (continuous, categorical, etc.) with the advantage of having in-built error estimates when used for applications that require data imputation. These RF methods also have the advantage of being applicable without the need for a test dataset (Stekhoven and Bühlmann, 2012). A broad range of non-parametric imputation strategies are widely discussed in the literature (Paddock, 2002, Roy et al., 2017).

A disadvantage of heuristic imputation methods is that they require expertise and experience to use effectively as their technical limitations of the learning algorithms are not always established (Paddock, 2002). Heuristic methods tend to consist of a number of not-readily-accessible evolutionary stages, potentially leading to systematic errors and cognitive bias. Further, since heuristic methods do not define parameter models of the data, they are less able in use for prediction. This is predicated on the fact that heuristic methods tend to rely on internal validation using a training and test dataset. External validation would require quantifying the confidence in the prediction, by being able to ascertain the results derived from scenarios in difference to those presented by both the training and test analysis data (Harrell et al., 1996). For urban mobility, in situations where a better understanding of the movement patterns is sought, an optimal solution resulting from an exhaustive search of the solution space is desirable, leading to a preference for parametric imputation and analysis methods.

7.2.3 Parametric imputation

Parametric data imputation strategies are developed by first creating equations (the models) that relate sets of quantities as explicit functions of a number of independent variables. These variables are indexed by fixed values or probability distributions called parameters of the problem. The parameters are unique characteristics of the problem (or population) being analysed. The analyst assumes that such a model explains the data generation process (i.e. the mechanism through which the data was created), and this model forms the basis for imputing any missing data values. A model of the parameter is created based on an intuition about the form of the probability distribution that would adequately represent the distribution of the parameter, and the likely distribution of the variables of the problem based on the data they contain. Over the years (Hald, 2008) standard forms of these models (equations) have been developed and established, empirically or theoretically, and found to be suitable for particular problems and scenarios. Parametric imputations (Allison, 2001) are principled methods as they do not assume knowledge of the missing values, but use information and the context of the observed data as the basis for estimating the missing values. In practise, parametric methods variously employ statistical tools, Bayes theories, stochastic and iterative strategies in the processes of deriving a solution to the chosen methodology, and can be developed using the so called Frequentist or Bayesian paradigms[65] (Bayarri and Berger, 2004)

[65] In the Frequentist paradigm, for the case of a coin flipped with probability 'θ' of heads, if we flip the coin 14 times and it ends up tails 4 times, then the best (maximum likelihood) estimate of 'θ' is $10/14$, which is a fixed quantity. Using standard notation, the probability of throwing heads H two times in succession is $Pr\{HH|data\} = (10/14)^2 = 0.51$. On the other hand, the Bayesian paradigm is based on the Bayes rule $Pr\{\theta|data\} \propto \Pr\{data|\theta\} * Pr\{\theta\}$ In essence, the posterior (distribution) of θ given the data is proportional to the product of the likelihood of θ given the data and our prior beliefs about θ. The toss of a coin is a binomial distribution, so the likelihood of θ is the binomial function. The posterior probability of throwing heads H two times in succession $Pr\{HH|data\}$ is a binomial distribution tuned by our prior beliefs. If our prior belief is that any value of θ is possible, then it can be shown that in the Bayesian framework $Pr\{HH|data\} = 0.49$. If our prior belief is that any of the means of θ is possible, then $Pr\{HH|data\} = 0.51$, which is equivalent in value to the Frequentist case.

A number of these principled parametric imputation methods have evolved in the literature (King et al., 2001, Schafer, 1997, Van Buuren, 2007), and the widely reported of these methods include the maximum likelihood (ML), multiple imputations (MI), and Bayesian imputation (BI). Implementations of these imputation methods in current available software (Buuren and Groothuis-Oudshoorn, 2011, Graham et al., 1996, Rubin, 1978) assume that the missing values are missing at random (MAR). The MAR assumption holds when the observed data values are sufficient to explain the missing data being imputed. In essence, the missing values are necessarily dependent only on the observed data values (see Chapter 2 and Chapter 3 for details), and the mechanism that caused the missing data is distinct from the analysis mechanism being inferred. Whilst the ML methods are only tenable when the data are MAR, the MI and BI methods can be developed by including missing data models to accommodate scenario when the data assumption of MAR does not hold (i.e. when the data are MNAR[66]).

The Maximum Likelihood methods (Allison, 2012, Yuan and Bentler, 2000) infer the missing data from likelihood functions of the observed data. To explain this in simple terms using standard notation, Bayes rule (see also footnote 34) in terms of the likelihood can be written as $Pr\{\theta|data\} \propto \mathcal{L}\{\theta|data\} * Pr\{\theta\},$ where θ is the model parameter, and $\mathcal{L}$ is the notation for likelihood. In principle, modelling aims to find the value of the parameter θ that best fits the data, and this is equivalent to ascertaining the value of the parameter (θ) that maximizes the data likelihood $\mathcal{L}$. When there are no missing values, and say M data variables with N observations each, the likelihood function is the product of the joint probability function $\mathcal{F}_i(..)$ of each observation i, and θ are the model parameters being estimated. This is written in Equation 7 (Allison, 2012). In essence, $\mathcal{F}_i(..)$ can be seen as the contribution of observation $'i'$ to the likelihood. If now there are missing values in the first two data variables, say N_{comp} complete observations, and N less N_{comp} observations with missing values in those first two variables. Then, the likelihood function[67] becomes that shown in Equation 8, where $\mathcal{F}_i$* is the joint probability density in the absence of the missing data values in the two variables.

[66] A detailed explanation of missing data types have been presented in Chapter 2 and Chapter 3.

[67] The likelihood function is deduced in the assumption that the missing data are MAR.

$$\mathcal{L} = \prod_{i=1}^{N} \mathcal{F}_i(data_{i1}, data_{i2}, \dots, data_{iM}, \theta) \tag{7}$$

$$\mathcal{L} = \prod_{i=1}^{N_{comp}} \mathcal{F}_i(data_{i1}, data_{i2}, \dots, data_{iM}, \theta) \prod_{i=(N_{comp}+1)}^{N} \mathcal{F}_i{}^*(data_{i3}, \dots, data_{iM}, \theta) \tag{8}$$

The $\mathcal{F}_i$* is the probability of observing the N_{comp} complete observations, deduced by integrating out the missing values. For brevity of the expression, assuming we replace $data_{ij}\ by\ t_{ij}$ the $\mathcal{F}_i$* is derived from $\mathcal{F}_i{}^*(t_{i3}, \dots, t_{iM}, \theta) = \iint \mathcal{F}_i(t_{i1}, t_{i2}, \dots, t_{iM}, \theta)\, dt_{i1} dt_{i2}$. In the case of missing data values, the Maximum Likelihood (ML) imputation finds the value of θ that maximizes the expression in Equation 8, using any of the range of numerical algorithms (Dempster et al., 1977).

The ML imputation has limitations in its use if the missing values are in the dependent variable, as then the likelihood expression reduces to the complete case scenario whereby observations with missing values are deleted (Allison, 2012). Another limitation of the ML imputation is that even when the missing values occur in the predictor variables, there are limited implementations of the ML imputation in current commercially available software. An ML imputation contrivance developed in the literature is a two-stage expectation maximisation (EM) and Monte Carlo process (Honaker et al., 2011). The EM is an algorithm for estimating the means, variances, and covariance of a design matrix, as the model parameters are functions of these. A regression exercise solved by a Monte Carlo procedure is then used to estimates the missing values. An example of the ML imputation procedure when the variables are assumed to have a multivariate normal distribution are available in the literature, as these highlight particular details of the ML imputation (Allison, 2012, Honaker and King, 2010).

Multiple imputation (MI) methods have wider application, and are developing prominence in the literature (Azur et al., 2011, Horton and Lipsitz, 2001, Rubin, 1996). The MI methods are typically a 3-stage process. First, the missing values (omissions) in the dataset are filled in with plausible values. These plausible values could typically be derived by an ad-hoc mean imputation. Second, a random error term is added to the regression expression created from the design matrix, whilst repeated regression imputations are performed whereby each repeat creates a set of plausible values for the missing data. Third, efficient inferences are created using complete data methods using the sets of plausible values (Allison, 2001, Ibrahim and Molenberghs, 2009, Royston, 2005). Typically, only about 5~10 repeats are recommended in the literature to produce estimates that are consistent, asymptotically efficient and asymptotically normal (Azur et al., 2011, Buuren and Groothuis-Oudshoorn, 2011). A general purpose MI method is the multiple imputation chained equations (MICE) method (Azur et al., 2011, Buuren and Groothuis-Oudshoorn, 2011, White et al., 2011) which is also a three stage process like the MI, but the second stage is an iterative process whereby after the missing data are randomly imputed, a new parameter value is drawn from the observed-data posterior distribution. This parameter value is used to impute the missing values, and then new parameter values estimated. This process is repeated thereby iteratively improving the estimates of the missing values. The process is akin to a Gibbs sampling procedure (Plummer, 2003, Smith and Roberts, 1993), and in implementing MI or MICE, the convergence of the iterative second stage needs to be assessed, as well as statistically evaluating the quality and confidence in the imputation results. The main drawback of the MICE is that it has no formal theoretical justification (Marchenko, 2011, Rubin, 1996), and is not geared toward accurately predicting the missing values, but result in valid statistical inference in scenarios of missing data. In many applications as in railway passenger mobility, the essence is to predict the missing values accurately as these for instance reflect in the estimates for journey factors for season tickets, and this in turn affects the comprehensive volumes estimated for passenger demand.

Bayesian imputations in the standard form are premised on the data being MAR just like ML, MI, and MICE imputation. However, Bayesian imputation can be easily adapted for application to scenarios where the data are MNAR i.e. missing-not-at-random (see Chapter 2 and Chapter 3 for detailed discussion of MNAR). This adaptation is achieved through the inclusion of a model to represent the statistical mechanism of the missing data (Lunn et al., 2012). Further, the Bayesian imputation is unique as the method simultaneously infers missing values as part of the system identification process. The Bayesian imputation models are also suited to hierarchical phenomena and spatial non-stationarity, which are particular features of mobility interaction (Brunsdon et al., 1996, Van Nes, 2002).

The Bayesian imputation is similar in concept to the procedure for MICE in the sense that the parameter estimates are iteratively improved through an MCMC based Gibbs sampling routine. In the Bayesian framework however, alternative sampling routines can similarly be adopted to optimize the iterative convergence of the parameter estimation procedure. The basic Bayesian imputation strategy assumes that the data are MAR, and that the parameters responsible for the missing mechanism $(\emptyset)$ are distinct from factors (θ) of the parametric model. In the case of LENNON ticketing data, θ and $\emptyset$ are distinct (Rubin, 1976) since $\emptyset$ is a design feature of the ticket measuring system, whilst θ is a property of the mobility dynamics.

In a fully[68] Bayesian framework an original covariate model is first created which regresses the outcome variable on all the other covariates. Then, additional models are created using one of the partially observed covariates as the outcome. This sequence is repeated for all the other partially observed covariates. As such, if there are M covariates with missing data values, then M sequential additional covariate models are created. The original outcome variable is not included in the additional sequence of models created, as the influence of the original outcome on missing covariate is implicitly included in the original model. Also, if multiple covariates have missing values, then once an additional model is created for a particular missing variable, that variable is not further included in any subsequent additional model. An additional feature of the Bayesian imputation is the creation of a missing data model which incorporates any additional knowledge the analyst might have about the mechanism of the missing data. If for instance in imputing the journey factors (i.e. daily rate of ticket use) that experience shows that longer distance journeys due to expedience are less likely to be made many times in the day, then this additional knowledge can be incorporated. The incorporation can be achieved perhaps by empirically determining how often a season ticket from say London to Manchester is used when compared to a similar shorter distance season ticket from say Bradford to Leeds. An missing data model may then indicate that the journey factors for season tickets are a function of the travel distance variable within the analysis dataset. The details of these such procedures are highlighted in the methodology within this chapter.

[68] The adjective ‘fully’ is used her to indicate that the analyst makes the choice of prior in the Bayesian model, and a sensitivity analysis is conducted to decide whether the solution is influenced by choice of priors.

7.2.4 Remarks

In this research, the Bayesian imputation framework is adopted because these methods increasingly have strengths previously only attributable to heuristic methods, and produce models suited for predictive analysis. LENNON ticketing data are supplied as cross-sections of tickets sold in thirteen 4-week periods per annum, forming a time-series cross-section (TS-CS) dataset.

In the research presented in this thesis, only one cross-section (4-week period) is analysed at a time. Within each cross-section of data, it is assumed that there are no sharp jumps in mobility activity. This assumption is premised on that in the rail industry period there are no sharp endogenous railway infrastructure changes like time-table alterations, new routes, lengthy line closures etc., and on this basis no sharp changes in mobility trends occur. In the rail sector, changes like time-table adjustments are only conducted after a period and typically during fare rounds. As a result, in the imputation analysis, the data has been assumed to be identically distributed within a period.

Hitherto, non-parametric heuristic methods would be adopted for the analysis of such TS-CS data to accommodate time trends in the mobility variables. In addition, semi-parametric methods have been used to introduce heuristics which accommodate space-time variations to capture non-linearity’s in TS-CS data by employing machine learning strategies (Shah et al., 2014). More recently however, Bayesian methods are now applied to analyse time variations as seen in publications in the literature (Brodersen et al., 2015, Dominici et al., 2002, Honaker and King, 2010). Although the TS-CS nature of the ticketing data is not fully exploited due to time limitations and expedience on the research project, such analysis of TS-CS ticketing data is discussed as work anticipated to be conducted for the future. Urban mobility tends to be a local-area phenomena, and Bayesian models are increasingly suited to analyse such hierarchical and spatially non-stationary mechanisms (Carlin et al., 1992, Gelman and Hill, 2006, Van Nes, 2002). Bayesian models are suited for the analysis of hierarchical, non-linearity, and complex phenomena like mobility on the railways. An ability to exploit such a framework while imputing missing attributes data is a particular advantage of the Bayesian modelling framework. Bayesian models are also suited to model interaction phenomena which included Poisson and negative Binomial distributions used to model variously dispersed count data associated with urban mobility (Lunn et al., 2012).

7.3 Technicality of LENNON data

Most analysis in the railways sector of Great Britain currently is based on ticketing data from LENNON (ATOC, 2003a). The LENNON database consists of over 2.7m daily ticket transactions sold from over 8000 ticket machines in stations, on trains and over the web. The revenue from these tickets are polled, validated and apportioned by the Rail Settlement Plan (RSP) to about 30 associated train operating companies (TOCs). Prior to 2008 an origin-destination matrix (ODM) was created direct from the LENNON ticketing data and then infilled with Travelcards and Airport links tickets sold by Transport for London (TfL) (ORR, 2014b). The ODM is a contingency table relating origin and destination train stations, and made up of journeys made per ticket, resulting revenue and passenger travel distance (ORR, 2012a). Since 2008 an industry owned replacement demand matrix called MOIRA[69] was created from LENNON augmented by TfL Travelcards, airport links, and tickets sold in major urban areas by Passenger Transport Executives (PTEs)[70]. MOIRA is now used to create ODMs. The PTE's (also called PTA's[71]) within the UK are shown in Figure 7.1. The Office for Rail Regulation's (ORR[72]) publishes annual ODM statistics on rail passenger travel in England, Wales and Scotland. There are known shortfalls in the current ODMs resulting from the shortfalls in the LENNON ticketing data. These shortfalls highlighted in Chapter 2 and Chapter 3 include missing information on passenger daily trip rates (journey factors) for season tickets, accurate tenures of flexible tickets, and precise flows to grouped stations.

[69] Recall that MOIRA is the acronym for 'model of inter-regional activity' which is the railway industry reference demand forecasting software to assess impacts of supply service (time-table) changes on passenger demand, enabling strategic planning.

[70] Passenger Transport Executives (PTEs) sponsored ticket sales in major urban areas outside Greater London. Transport ticket products sold in the various PTE zones tend to be unique to that zone; as such mobility analysis for a particular PTE would typically be tailored. About 28 Train Operators (TOC's) manage trains and stations across regions of Britain.

[71] The PTA's are the Passenger Transport Authorities, now renamed Integrated Transport Authorities (ITAs) since the Local Transport Act of 2008.

[72] The Office for Rail Regulation's (ORR), now called Office of Rail and Road (ORR) since 1st April 2015.

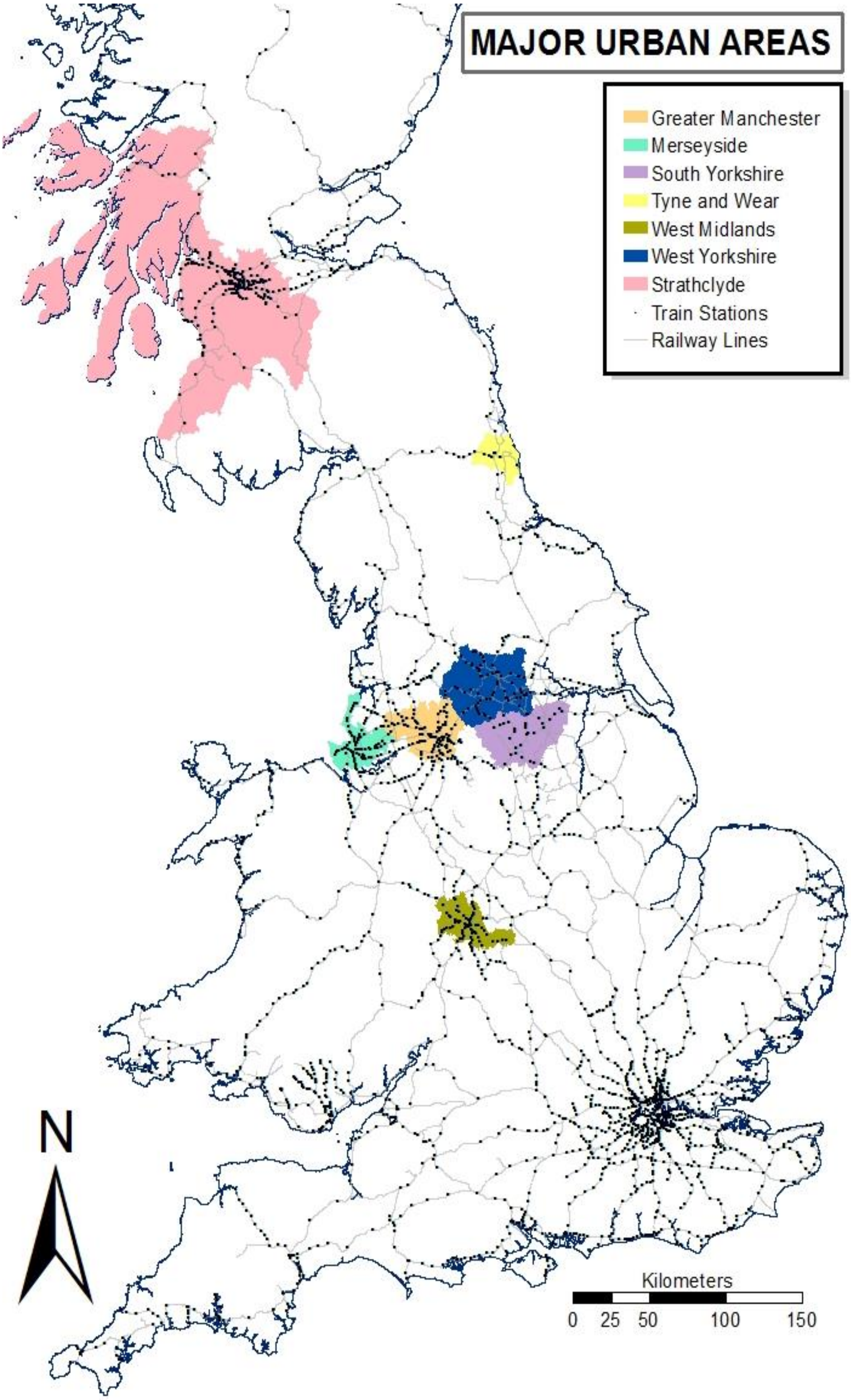


Figure 7.1 Major PTE Urban Areas in UK (excluding Greater London).

The current industry strategy for resolving the missing data gaps in LENNON are to use proxy trip rates for the journey factors. These proxies are empirical estimates from dated surveys. Similar proxies are developed for flexible tenure tickets. Flows to group stations and any unknown direction of passenger travel are resolved by reconciling flow estimates with those of dated surveys like the National Rail Travel survey (NRTS), and then randomly assigning ticket flows to achieve those estimates from the NRTS. Infilling and adjustments for PTE tickets are based on ad-hoc manual fixes. For example, MetroCard Zone 1-5+6 tickets (sold by regional operators) are unlikely to be used on a Harrogate-Leeds journey, as the cheaper Rail Settlement Plan (RSP) tickets (sold by UK National Rail) would be the more likely ticket choice. Various private companies (SDG, 2011, SDG, 2012, SDG, 2016, Taylor, 2013b) had severally been commissioned in collaboration with rail industry working groups (ORR, 2014a, ORR, 2012b) to produce improvements to MOIRA, but the shortfalls persist and have impacted on the consistency and accuracy of annual flows reported (ORR, 2014c). This calls for a review and validation in line with current scientific best practices of the infilling and imputation strategies used to augment LENNON and produce MOIRA. This forms the crux of work presented in this chapter.

As mentioned earlier (in Section 2.3), geographic datasets arranged in rectangular attribute tables consist of records (rows) representing tuples and fields (columns) representing variables. As introduced earlier, in the LENNON table each row in the table represents a ticket sale to an individual (or to groups who purchased ticket products with the same description). Each ticket includes attributes like purchase time (monthly period in the case of the anonymized dataset supplied to the project), station entry (origin) and station exit (destination), estimated journey factor (number of journeys per ticket), estimated track miles covered per ticket and cost, ticket type and additional product codes describing further details about the ticket. Missing data can occur at a primary level when one or more complete records (rows) are missing and this is called unit omission. At a secondary level when values for one or more variables within a unit are missing it is then called item omission. The scenario of unit omission is beyond the infilling scope of this thesis as it is assumed that all tickets sales are known, but that some ticket details (journey factors for season tickets, flows to precise grouped stations or PTE stations etc.) necessary for accurate forecasts of passenger demand are missing, referring to the scenario as item omission.

The missing journey factors for season tickets have historically been estimated using ad-hoc methods (ORR, 2012b, ORR, 2014c), resulting in the fractional journeys highlighted in the snippet LENNON data in Table 3.1 (Section 3.2.2). Another example of an item omission within the West Yorkshire ticketing data are entry and exit stations described as Bradford BR, Pontefract BR, Wakefield BR, etc., with the BR (British Rail) suffix. The train stations with BR name extensions do not exist on the rail network, but within the LENNON database they represent a group of stations, with precise flows to individual stations within the group unknown (missing). An illustration in Figure 7.2 shows flows from say Leeds to Bradford Forster Square, as distinct from flows from Leeds to Bradford Interchange. Whereas the LENNON ticketing database prior to validation many-a-time would not distinguish these two flows, but describe them as both from Leeds to Bradford BR.

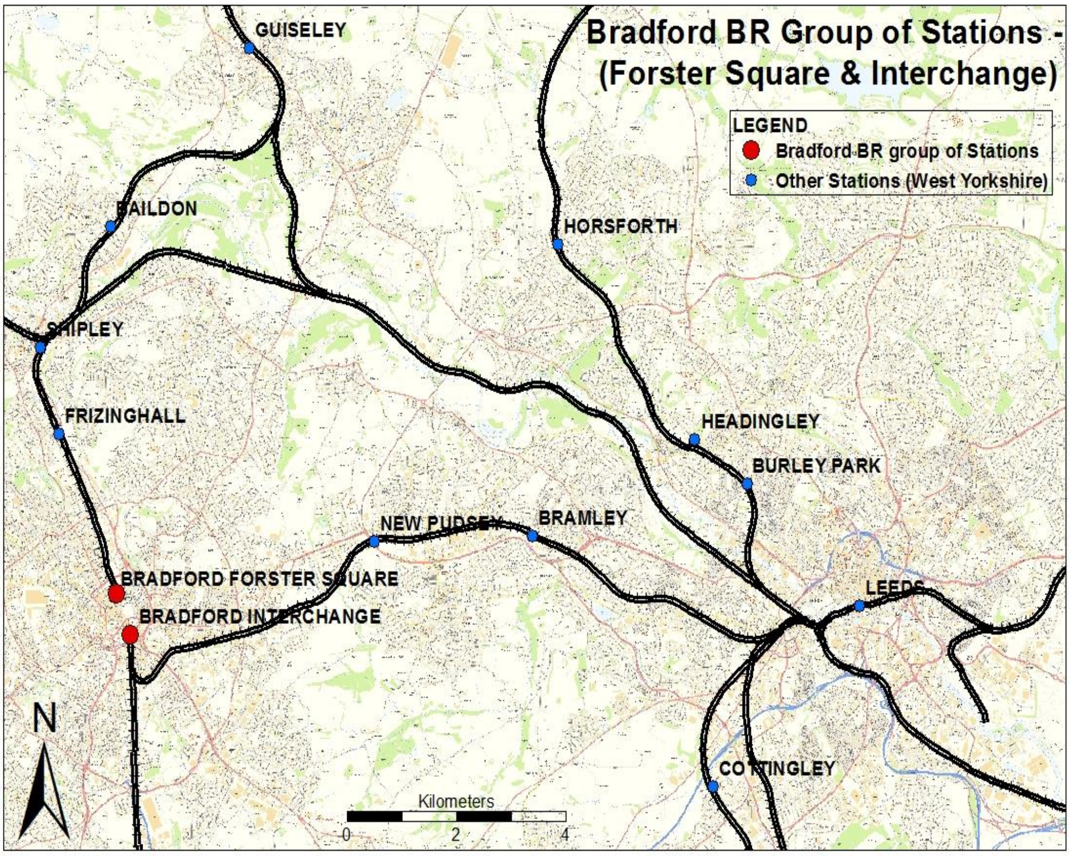


Figure 7.2 Bradford (BR) Group[73] of train stations.

[73] Typically in the LENNON dataset, it is observed that lengthy journeys to railway stations located in close proximity and serving overlapping regions are ascribed to a group name. However, shorter journeys to the same train stations would be assigned tickets indicating the specific

The Bradford Interchange and Bradford Forster Square stations are in close proximity (~700m) but the mobility flows to these stations are via completely different routes, and when precise flows to these stations are not recorded within LENNON, the flows are unknown (unobserved) and structurally classed as an item omission. Such missing information on precise flows to group stations are relevant in creating a rich picture of mobility, and is necessary for causal inference on drivers of mobility on the rail network. Item omissions can further be sub-divided into three patterns: univariate, monotone, or random.

Univariate omission is where all the missing values occur within only one single variable. An example would be if the journey factors for season tickets were the only missing data values in the LENNON datasets. Then such an omission would be classed as univariate as it only occurs in the variable for journey factors. Monotone omission is where the missing values occur systematically across the set of variables being investigated. An example would be when say regional PTE MetroCard tickets are simply appended to LENNON tickets from RSP. Then, if the PTE tickets consistently do not have precise information on specific entry and exit stations, journey factors and distances covered, such missing information which would be systematic (across the PTE tickets) would make the entire dataset to be described as having a monotone pattern of missing values. In some instances, a monotone pattern of the missing values is not readily apparent from visual inspection of the data table. However, there exists software that enables an assessment of a data table for monotone patterns by re-organizing the data variables and tuples to see if any monotone patterns are exposed. There exists special imputation algorithms particularly suited to monotone patterns of missing values (Horton and Lipsitz, 2001, Van Buuren, 2007, Yuan, 2010), so an identification of such patterns would enable an exploitation of such specialist imputation methodologies. A random pattern of omission is the case where the missing values are randomly scattered across the data table with no obvious deducible systematic patterns of missing values. A classic random pattern of missing data values would be a randomized control trial where all the measurements that were not captured during the survey would be classed as missing, with a random pattern of missing values.

name of the entry and exit station. There is no specific indication of the length of journey that warrants a flow to be assigned to a group station, and perhaps the criteria for assignment is a historical nuance in the railway industry.

7.4 Method for Bayesian imputation

The methodology for Bayesian modelling and imputation is presented in this section. The main regression model is presented containing the outcome variable and covariates. Some of these covariates have missing values, so other relevant covariates created from m-IPF and GIS-GTFS are selected (see Little’s Test in Section 3.2.3), such that the combined set of variables are sufficient to explain the missing data. As mentioned earlier, in the Bayesian framework separate data models are created for all the covariates with missing values. These covariates are the journey factor, ticket validity period, and group station flows. For each model, directed acyclic graphs (DAGs) are created (Pearl et al., 2016). DAGs are directed graphs used to visually portray the crystallisation of variables in the data generation sequence. DAGs conceptualize the relation between the covariates, and provide a systematic approach to modelling consumer and observational data, as a basis for subsequent objective causal inference (Pearl et al., 2016). Once the DAGs are created and reconciled with the regression and imputation models, the BUGS Bayesian software (Lunn et al., 2000, Plummer, 2003, Spiegelhalter et al., 2007) is used to identify the parameters[74] of the model. Finally, the regression results for the main model are presented, alongside the imputation results for those covariates with missing values. These identified missing values (journey factor, ticket validity period, and group station flows) from Bayesian modelling are the behavioural passenger attributes which when infilled into LENNON further enhance the ticketing data. The LENNON ticketing data having been enhanced by m-IPF, GIS-GTFS, and Bayesian imputation, form a requisite dataset for subsequent mobility analysis as presented in the next chapter (Chapter 8).

7.4.1 The Bayesian set up

In the Bayesian framework, modelling and imputation occur simultaneously. Whilst the modelling is where the conventional parameters of the regression model are identified, the imputation is the identification of values for any missing covariate data items. The Bayesian regression model is similar to classical regression in that the model of the sampling distribution of the data is specified, and then a relationship is formed between that distribution of the response variable and the explanatory variables. The difference however, is

[74] Note that in the Bayesian framework, parameters refer to both the traditional system parameters as well as the missing values.

that in the Bayesian modelling framework, prior distributions are specified for the regression coefficients and unknown parameter values (Gelman et al., 2014b, Lunn et al., 2000, Plummer, 2003), instead of assuming that they are constant as in the classical frequentist regression (Best et al., 2000, Ewing, 1974, Gardner et al., 1995, Jaccard and Turrisi, 2003). The Bayesian approach is illustrated in Equations 8 where in Equation 8a the response variable Y_i is count data modelled as a negative binomial distribution. Equation 8b is typical of a regression model in the classic frequentist approach, where X_k represents the k^{th} covariate with corresponding coefficient β_k. The inclusion of Equation 8c is the difference between Bayesian and classical regression modelling. Equation 8c makes provision for specifying the priors (currently left blank). The choice of priors is a pertinent consideration typically developed during model implementation in Bayesian modelling. The choice of priors could be informative, vague or non-informative, improper or some mixture of these. As a general rule, if a sensitivity analysis indicates that the choice of prior is influential on the results, then robust conclusions cannot be drawn from the data model. Under such conditions a more informative prior (based on say industry knowledge of the particular problem) or an alternative model should better inform the prior.

$$Y_i \sim Negative\ Binomial(\rho, \gamma) \tag{8a}$$

$$log(\mu) = \beta_0 + \sum_{k=1}^{p} \beta_k X_k \quad \text{where} \quad \mu = \gamma((1-\rho)/\rho) \tag{8b}$$

$$\text{with priors} \quad \beta_0 \sim dflat(\,); \quad \beta_k \sim Normal(\,,\,); \quad \gamma \sim Gamma(\,,\,) \tag{8c}$$

In Bayesian imputation, the missing values in any of the covariates are simply left as null (NA), and during the simultaneous process of solving the regression model, any missing values are inferred from the posterior, just like the unknown model parameters would be identified. During the solution an appropriate sampler is selected based on the composition of the model, such that posterior samples would converge to the appropriate posterior distribution. As with many principled imputation strategies, the Bayesian framework implicitly assumes that any missing data values in the covariates are missing at random (MAR). In this research, considerable effort is made (through the m-IPF and GIS-GTFS applied to LENNON ticketing data) to achieve a dataset that is MAR. However as alluded to earlier, a strength of the Bayesian framework is that it includes the possibility of a missing data model to ensure that the imputed values are a true reflection of the mobility mechanism being modelled. As such, a main mobility model is created to reflect the relationship between the count of passengers and the other covariates (Equation 8). This mobility model is then complemented by missing data models for the journey factor, ticket validity period and flows to group stations (the details are presented below).

For prediction and imputation, the comprehensiveness of a mobility model is a desired feature (Gelman, 2007, Knutti and Sedlácek, 2013), and in such exercises, as has been done in this research, the relevant covariates derived from the spatial microsimulation (m-IPF) and network simulation (GIS-GTFS) are included. This ensures the inclusion of the heterogeneous effects of all the variables in the mobility phenomena recorded. For causality and interventions case studies however, there is a need to be more parsimonious by creating models that capture the essence of mobility while simplifying its features (Angrist et al., 1996, Ho et al., 2007, Pearl, 2009, Shadish et al., 2002). The detailed model specification with the covariates designed for imputation and prediction are presented in the following sections within this chapter, along with the directed acyclic graphs DAG[75] (Pearl et al., 2016). The detailed list and names of the exogenous and endogenous covariates used to build the Bayesian imputation model can be found in the scripts in Appendix D, included in the CD accompanying this thesis.

7.4.2 Main mobility model

The regression mobility model of interest is designed to exploit the range of modelling capabilities within the BUGS Bayesian modelling framework. The model constructed is described as a negative Binomial, centred, spatial-interaction, non-linear, hierarchical model shown in Equation 9. The model is negative Binomial because the outcome variable is the count of passengers on a particular origin destination (OD) flow on the rail network. Due to the aggregation of the LENNON ticketing data supplied to the project, the count is made up of passengers who have purchased similar tickets products, so that the counts are disaggregated by detailed specification of ticket product.

[75] A directed acyclic graph (DAG) models probabilities, connectivity, and causality between variables. DAGs use nodes denoted by circles and edges connecting the nodes as primitives. The acyclic nature of DAGs implies a trace from any first node to other nodes will not re-visit the same node twice. This feature of DAGs makes them particularly useful for probabilistic causal inference and intervention studies

As widely established and reported in the literature, the likely distribution of count data are typically Poisson or negative Binomial (Gardner et al., 1995). The choice of negative Binomial has evolved during the project as such a model was seen to be less sensitive to the choice of parameter priors, and yielded MCMC chains that converge efficiently. The choice of likelihood function is typically part of the model developmental stages. However, negative Binomial distributions are known in the literature to be better than Poisson distribution when modelling count data with high variability (over-dispersed) (Gschlößl and Czado, 2008, Rodrıguez, 2013), non-uniform variances (i.e. heteroscedastic), and with many zeros in the design matrix, i.e. zero-inflated (Gardner et al., 1995, Greene, 1994)).

The model is also described as centred because all the continuous covariates are each normalised by subtracting a value (equivalent to the mean of the particular variable) from every value of the particular variable. The practical advantage found in centring is that the MCMC chains developed to identify the parameters and missing values converge more efficiently, and with narrower credible intervals, indicative of a more precise solution. However, as reported in the literature (Lunn et al., 2012), centring is useful in reducing the correlation between parameters of the model, thereby enabling the MCMC algorithm to explore the parameter space more efficiently (Albert, 2009, Kruschke, 2014, Lunn et al., 2012).

Further, the main mobility model is described as spatial-interaction nonlinear, because the entry and exit stations are included categorical covariates of the model. These entry and exit points reflect the mobility interaction of passengers in space. In addition, the main model includes the logarithm of a variable derived from the along-track distance travelled by passengers. The non-linearity is introduced by taking the log of the distance, and the parameter of this variable represent a propensity of passengers to resist travelling longer distances between an origin and destination. Finally, the model is described as hierarchical as a distinction is made between passengers who use one-way, return, and flexible season tickets. The assumption is that passengers on flexible tickets behave quite distinct from those on inflexible tickets. By creating a hierarchical model, all passengers belong to the same hyper group of railway passengers, but belong to separate sub-groups depending on the particular ticket type that the passenger uses. The notion of hierarchical models is founded on the concept of exchangeability and conditional dependence (De Finetti, 1972).

This main model in Equation 9 was alluded to in presenting Equation 8 earlier. The Equation 9 is structured similar to Equation 8, but is reproduced more elaborately in Equation 9, with the specification of the priors excluded. The specification of priors is part of an elaborate endeavour in Bayesian modelling, and typically a iterative development exercise. This exercise involves assessing both the sensitivity of the model results to the choice of prior, as well as the effects of the model initial conditions on the convergence of the model.

The outcome variable $Y[i]$ represents the i^{th} passenger count for journeys from station entry $[O_i]$ to exit $[D_i]$. Each passenger in the count $Y[i]$ made a journey with an i^{th} type of ticket that was used J times per day. Note that the ticket-type is a categorical variable, so that for instance the i^{th} ticket and the k^{th} ticket within the ticket-type variable could be the same. Each of these tickets have a validity period V. The journey factors and ticket validity (J and V) are rates and as such are operated on by the log (i.e. offset) as indicated in Equation 9b. Subsequently, J and V can be transferred to the right of Equation 9b as is the practise in dealing with rates and offsets in classic regression models (Frome, 1983, Zeileis et al., 2008). There are missing values in the covariate for journey factors (J) and in the covariate for ticket validity period (V). The Bayesian framework treats any missing observations as parameters of the model, just like the coefficients β which are yet to be identified and as such unknown. In the BUGS implementation of the Bayesian modelling (Lunn et al., 2012), the solutions are implemented by optimizing the choice of sampling technique to adopt (i.e. Gibbs sampling, Metropolis-Hastings, Rejection sampling, etc.) in the MCMC algorithm.

$$\boldsymbol{Y}[\boldsymbol{i}] \sim \boldsymbol{Neg.Binomial}(\boldsymbol{\rho_i}, \boldsymbol{\gamma}) \qquad \text{where} \qquad \boldsymbol{\mu_i} = \boldsymbol{\gamma} * (\boldsymbol{1} - \boldsymbol{\rho_i})/\boldsymbol{\rho_i} \tag{9a}$$

$$\text{then} \quad \boldsymbol{log}\left(\frac{\boldsymbol{\mu_i}}{\boldsymbol{J_i} * \boldsymbol{V_i}}\right) = \boldsymbol{\beta_o}[\boldsymbol{O_i}] + \boldsymbol{\beta_d}[\boldsymbol{D_i}] + \sum_{\boldsymbol{m}=\boldsymbol{1}}^{\boldsymbol{\varpi}} \boldsymbol{\beta_{mis}}\,[\boldsymbol{M}[\boldsymbol{i}]] + \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{v}} \boldsymbol{\beta_k}\,[\boldsymbol{X_{ki}}] + \boldsymbol{\beta_{OD}}[D_{\boldsymbol{OD}}] \tag{9b}$$

Note that the $[\boldsymbol{O_i}]$ and the $[\boldsymbol{D_i}]$ described in Equation 9b refers to the station entry and exit respectively, which are distinct from the passenger's residence of original and final work destination, which in other contexts can also be described as origin-destination. In the developing the mobility model, both the residence and entry station, as well as exit station and final destination were considered.

In traditional spatial interaction models $Y[i]$ would be structured as a matrix (Dennett, 2012), but in this instance the variables are left in the vector format due to the increased in number of variables associated with each $[OD]$ pair. These variables include ticket type, journey purpose etc. In essence, due to the increased variability in the datasets, a specific $[OD]$ pair could be traversed by passengers with different ticket types. As such it is not possible to aggregate into $[OD]$ pairs to sustain the traditional rectangular $[OD]$ interaction matrix. The vector $[OD]$ format thus stems from the need to exploit the variability in the ticketing data by distinguish the heterogeneity in passengers on the same $[OD]$ flow. The $\left[M[i]\right]$ in Equation 9 represents an internal nested indexing nomenclature used to represent categorical covariates (Lunn et al., 2012). The nested indexing precludes the need to re-model each categorical covariate into separate binary variables as is typical in regression (Long and Freese, 2006). As depicted in Equation 9, there are ϖ variables which are categorical in the model, and these are represented by $\left[M[i]\right]$ when associated with the i^{th} passenger count. For example, one of the categorical variables in the $\left[M[i]\right]$ is the group station indicator variable. There are missing values in that group station indicator covariate as described next.

Recall that during the data pre-processing, if a single flow occurred to a group of stations, we would not know which particular station within the group through which the flow actually occurred. If there are say 4 stations A, B, C, D in a group, and a ticket indicates a flow to this group of stations, then, there is no way of identifying whether the passenger actually travelled through station A, or B, or C, or D. The analysis procedure developed consists of augmenting that single flow by replacing it with four flows, one to A, one to B, one to C, and one to D. These would form a set of augmented flows, and would be identified as such in the consumer data table. However, each set of augmented flows are actually a single flow multiplied by the number of stations in the group. As the particular station indicated by the flow is not known, all the augmented flows in the Bayesian modelling framework are indicated as NA, with flows to non-group stations indicated as ‘1’. As such the condition imposed on the group station indicators is that the sum of the NA’s for augmented flows to the group stations would amounts to ‘1’. In essence $\sum_{i=m}^{n} I_i = 1$, where I is the group station indicator, m is the index of the first station in the group, and n is the index of the last station in the group sequence.

In Equation 9b, the $[X_{ki}]$ represents the continuous variables, and there are ν continuous variables. The $[D_{OD}]$ represents the non-linear negative exponential (or power law) distance decay which is used to model the propensity of passengers to resist travelling further distances to fulfill activities. The power law or negative exponential models of distance decay are established in the spatial interaction modelling literature (Dennett, 2012, Wilson, 1971).

In formulating the models, an assumption of missing at random (MAR[76]) is made premised on that the additional variables from the NRTS, Census, and GIS-GTFS network simulation are necessary and sufficient to explain the missing values in the LENNON datasets. As such, the '*missingness'*[77] of the enhanced rail data (created from spatial microsimulation of NRTS-Census-LENNON data and GIS-GTFS network simulation) can be described as MAR. To complement the development of a robust modelling framework, and to forestall a situation where the missing data might not be MAR as assumed, main mobility models are constructed to describe the fundamental passenger mobility interaction. In addition to these main models, additional missing data models are created for those covariates with missing data values. Further to these two models (i.e. main and missing data), models are constructed if knowledge is available of the statistical michanism (***missingness***) through which the missing values became unobserved. Any further mobility information available to the analyst are incorporated within this missingness model. Directed acyclic graphs (DAGs[78]) are further created to visually represent the conditional dependencies of the covariates used in the models (as subsequently explained).

[76] The missing at random (MAR) concept was discussed earlier in chapter 4 and refers to the situation where the observed values (X_{obs}) capture the entire essence of the system parameters θ, and are sufficient to explain the difference between the observed and missing data (X_{mis}). In such circumstances the missing data is uninformative, and the data MAR.

[77] The phrase '***missingness***' refers to the statistical mechanism of missing data, to distinguish whether the missing data is dependent on the magnitude and/or values of the observed data or those of the missing unobserved data. (Rubin, D.B. 1976).

[78] DAGS are the basis for solutions provided by Bayesian networks, and is a probabilistic graphical model used to represent the conditional dependencies of the covariates.

7.4.3 Missing covariates models

The methodologies for the development of the three imputation models are presented. The models enable infilling of missing values in covariates for journey factor, ticket validity periods, and flows to group stations. Aspects of the R-studio software codes used to implement the methodology are presented to enable the results created to be reproduced, and to highlight the strengths of the Bayesian framework. The details of the full R-scripts of the model specification used to implement the Bayesian imputation are described in Appendix D, and included in the CD accompanying the thesis.

7.4.3.1 Missing journey factors

The journey factor is the daily rate of ticket use. The LENNON information on season ticket products do not reveal how many times the ticket is used. As such for season tickets, the journey factors are required to be imputed to attain accurate passenger volumes and forecasts of demand on the rail network. Single and day return tickets are respectively used once and twice daily, so their journey factors are fixed at 1 and 2. Season tickets on the other hand have no restrictions on their daily rates of use, and can daily be used by the passenger an unlimited number of times. As a result, specific journey factors for season tickets are unknown (missing values).

In the Bayesian framework, once a main model describing the phenomena being modelled is created, additional imputation models can be included to complement this mode, also models are created to exploit any disparate information that might be available. The Bayesian framework is particularly useful in this regard by creating models that harvest data from different sources. Additionally, Bayesian BUGS language is declarative (as opposed to procedural traditional statistics packages), meaning that the order of presentation of the models is unimportant. These concepts are applied in creating imputation models for those covariates with missing items.

A set of chained models are used to impute the missing journey factors. In the Bayesian framework, these models are implemented in parallel, that way a feedback is achieved between the variables of each of the models. The journey factor imputation models are stated formally in Table 7.1 (with further details in Appendix D and the CD accompanying the thesis). This journey factor imputation model is made up of three chained models ran in parallel (see Table 7.1). Model 1 within Table 7.1 is the main mobility model relating the count of passengers (i.e. ‘Freq’) used as the outcome variable to the origin and destination zones, and a range of potential confounding covariates. This model is derived from Equation 9, and is a negative Binomial model. Model 2 in Table 7.1 is the actual journey factor imputation model for the missing journey factors values. In the Bayesian framework the outcome variable in the missing data model is the covariate with missing values, in this case the journey factors. Model 2 is a truncated Poisson regression model relating the journey factor rates to the passenger origins and destinations (as well as a range of potential confounding covariates). In the Bayesian framework, the count (i.e. the 'Freq' outcome variable) is excluded from the missing covariate model. Similarly, any subsequent missing data models will not include previously used outcome variables (Kruschke, 2014, Lunn et al., 2012). However, other explanatory variables can be included in the missing data model. The third Model 3 (within Table 7.1) is a specific ***missingness*** model for the journey factors. The addition of this sort of Model 3 enables the incorporation into the modelling process any additional specific knowledge about the missing journey factors. In this case the statistical mechanism through which the daily rates of ticket use (journey factors) became missing indicates that the propensity for high daily rates of ticket use increases with shorter travel distances. Model 3 is a Bernoulli model.

Further details of the models are included below, as Model1, Model 2, and Model 3 complement one another. It is noteworthy to point out that other modelling construct were considered in the research as part of the model development process. For example, for the main mobility model (Model 1), a Poisson regression was investigated as the outcome variable was count data. However, to accommodate the large variability in the rail mobility phenomena, the negative Binomial model yielded better Bayesian convergence characteristics. A mixed normal model was similarly explored for potential use as the actual journey factor imputation model (Model 2). The mixture of normal model could potentially reflect and model passenger use of restricted single and return tickets, as distinct from passenger use of the more flexible season tickets. A mixture of normal distributions could capture the categories of passengers who use the different types of tickets. However, the mixture of normal model negatively affected the efficient convergence characteristics of the main mobility model. As such, a mixture of normal model was not adopted for Model 2. The Bayesian modelling framework has potential to yield models that better reflect realistic phenomena; however at the current state-of-the–art of Bayesian modelling, considerable time is typically invested in developing such models.

Table 7.1 Formal model of the missing ‘journey factors’.

Outcome explanatory (O) or Exposure predictor (E) variable	Model 1: passenger flows	Model 2: journey factor	Model 3: ***missingness***
Passenger count (Freq)	O/NL		
Origin - station entry	E/N	E/N	
Destination – station exit	E/N	E/N	
Distance (journey, access, or egress)	E/N	E/N	E/U
Cost (monetary, or time)	E/N	E/N	
Daily Rates of ticket use – journey factor	E/N	O/PL/T	E/N
Ticket validity period (days/months)	E/N	E/N	
Group station flow	E/N	E/N	
Missingness indicator			O/BL

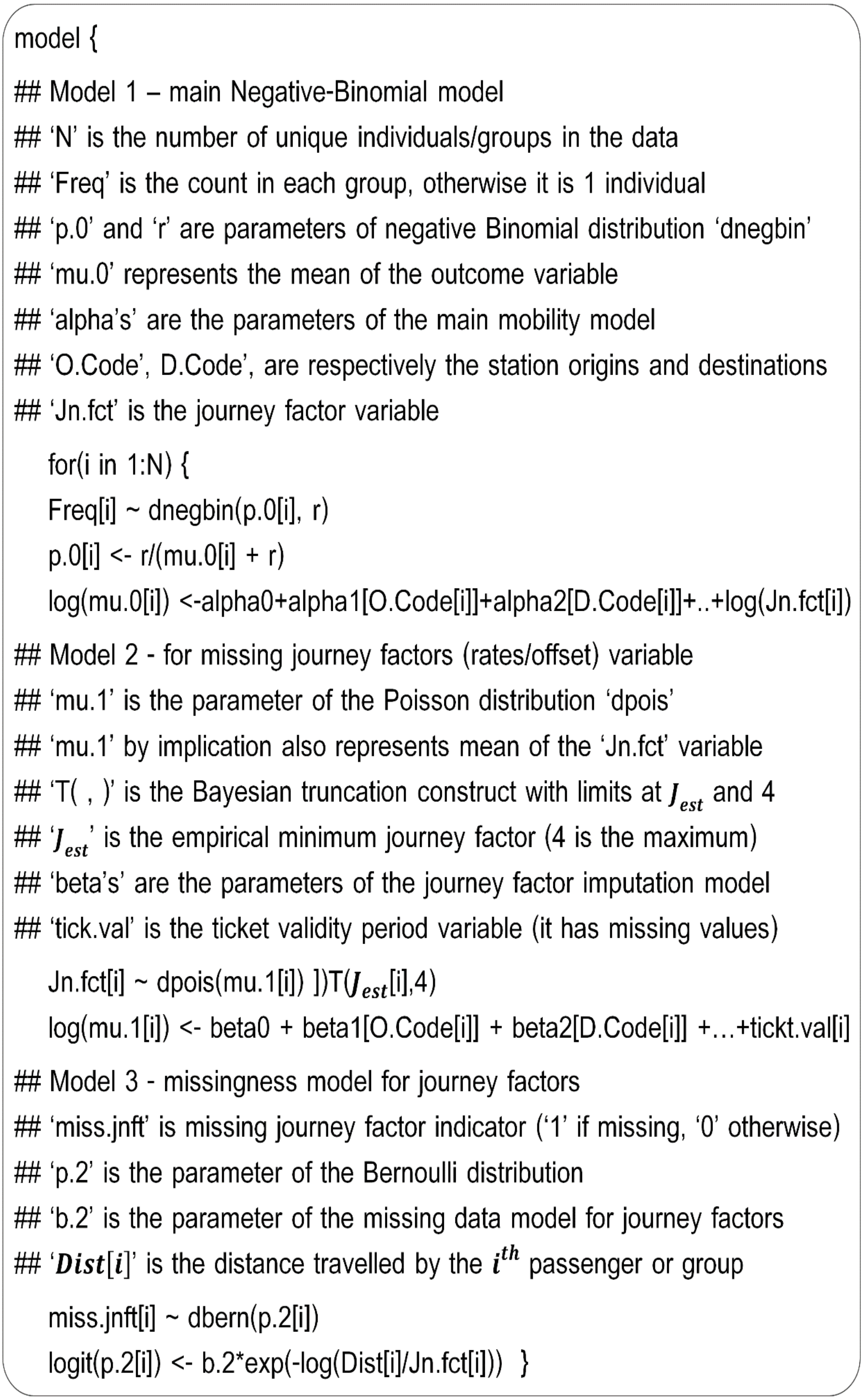
The type of outcome likelihood model (i.e. Poisson, negative Binomial, etc.) is indicated in the table above. The likelihood (L) for the journey factor imputation model can be either one the following: negative Binomial (N), Poisson (P), or Bernoulli (B). For example, in Model 1, the passenger count is the outcome variable, and its likelihood is modelled as a negative Binomial regression. As such, the passenger count (Freq) row for Model 1 would indicate O/LN (i.e. outcome variable as negative Binomial likelihood (NL). Had the likelihood been modelled as a Poisson regression, then the passenger count row for Model 1 would have been O/PL (with the PL indicating Poisson likelihood, while BL would indicate a Bernoulli likelihood model).

The prior specification for the exposure (E) variable can be modelled by a wide range of functions. In Table 7.1 the prior specifications adopted were normal (N), gamma (G), and uniform (U) distributions. For example, the distance exposure variable as used in Mode 2 has a normal distribution prior. As such, the distance variable in the column for Model 2 is indicated as E/N (i.e. and exposure variable (E) with a normal prior (N). It is noteworthy to point out that the parameters of a prior distribution (say the mean and variance parameters of a normal distribution) can in-turn have priors. In Bayesian models also, a lot of consideration is given to the choice of initial values (detailed in the appendix) to ensure that a stationary convergence is achieved efficiently (i.e. in reasonable time).

Bayesian models unlike frequentists enable constructs to be easily included in the modelling process. The Bayesian framework has a number of these constructs that enable incorporation of industry experience and knowledge in the imputation process to facilitate the construction of more realistic models. For instance, the missing journey factors for season tickets have been researched and estimates are available from industry experience (SDG, 2016). Industry estimates suggest that journey factors for season tickets lie on average between 1 and 4 times daily. This passenger behaviour derived from industry knowledge can be incorporated by truncating the journey factor to be between 1 and 4, using the $T(,)$ construct. This has been included in Model 2 in Figure 7.3. A more realistic truncation is achieved by implementing a more varied truncation limit dependent on the particular ticket product the passenger uses. This is captured by introducing a variable lower journey factor limit within the construct $T(J_{est}[i], 4)$, where $J_{est}[i]$ is the industry journey factor estimate for the type of ticket product associated with the i^{th} passenger or group. To reflect the limitation on the range of values that journey factor can take, the truncation construct used to ensure that the journey factor lies between say 1 to 4 dependent on the distance travelled is shown in Table 7.1. If for example an outcome variable is truncated (T) or censored (C), this is added to the appropriate variable. For example the journey factor in Model 2 is a truncated outcome variable indicated as O/P/T (with the T indicating a truncated variable.

A further missing data model can be added to reflect that the propensity to make repeated daily journeys reduces as the length of the journey increases. For instance, it is less likely that a long journey between London and Edinburgh will be made more frequently in a day, than a shorter journey from Leeds to Wakefield. This knowledge can be incorporated in the missing data model as shown by Model 3 within Figure 7.3 (see footnote[79]). In essence, the propensity for higher daily rates of use of season tickets decreases as the distance travelled with the ticket increases. In Model 3 (within Figure 7.3), a missing data indicator ($miss.jnft[i]$) is used to indicate if the journey factor is missing thus forming the basis for implementing the missing data model. For the Bayesian enthusiasts, the pseudo code specification of the model is included in Figure 7.3. Further detailed specification of this journey factor model is included in Appendix D and the CD accompanying the thesis.

[79] Model 3 was subsequently excluded from the analysis as in this instance the inclusion of an extra missing data model did not improve the solution.



```
model {
## Model 1 – main Negative-Binomial model
## ‘N’ is the number of unique individuals/groups in the data
## ‘Freq’ is the count in each group, otherwise it is 1 individual
## ‘p.0’ and ‘r’ are parameters of negative Binomial distribution ‘dnegbin’
## ‘mu.0’ represents the mean of the outcome variable
## ‘alpha’s’ are the parameters of the main mobility model
## ‘O.Code’, D.Code’, are respectively the station origins and destinations
## ‘Jn.fct’ is the journey factor variable
   for(i in 1:N) {
   Freq[i] ~ dnegbin(p.0[i], r)
   p.0[i] <- r/(mu.0[i] + r)
   log(mu.0[i]) <-alpha0+alpha1[O.Code[i]]+alpha2[D.Code[i]]+..+log(Jn.fct[i])
## Model 2 - for missing journey factors (rates/offset) variable
## ‘mu.1’ is the parameter of the Poisson distribution ‘dpois’
## ‘mu.1’ by implication also represents mean of the ‘Jn.fct’ variable
## ‘T( , )’ is the Bayesian truncation construct with limits at $J_{est}$ and 4
## ‘$J_{est}$’ is the empirical minimum journey factor (4 is the maximum)
## ‘beta’s’ are the parameters of the journey factor imputation model
## ‘tick.val’ is the ticket validity period variable (it has missing values)
   Jn.fct[i] ~ dpois(mu.1[i]) ])T($J_{est}$[i],4)
   log(mu.1[i]) <- beta0 + beta1[O.Code[i]] + beta2[D.Code[i]] +…+tickt.val[i]
## Model 3 - missingness model for journey factors
## ‘miss.jnft’ is missing journey factor indicator (‘1’ if missing, ‘0’ otherwise)
## ‘p.2’ is the parameter of the Bernoulli distribution
## ‘b.2’ is the parameter of the missing data model for journey factors
## ‘$Dist[i]$’ is the distance travelled by the $i^{th}$ passenger or group
   miss.jnft[i] ~ dbern(p.2[i])
   logit(p.2[i]) <- b.2*exp(-log(Dist[i]/Jn.fct[i]))  }
```

Figure 7.3 Missing ‘Journey Factors’ script.

The Directed Acyclic Graph (DAG) created to portray the relationship between the variables and covariates of the main model, the journey factor imputation model, and the missing data model is shown in Figure 7.4. The DAG serves to combine the individual declarative mobility and imputation models, and forms the basis of the MCMC algorithms implemented in the BUGS software. The DAG layout of Figure 7.4 is unconventional in two regards: first the covariates of the model are not explicitly specified, such that θ_F represents the combined parameters representing all the covariate parameters of the negative binomial model (Model 1). Similarly, θ_J represents the combined parameters of the missing journey factors model (Model 2), and θ_D represents the combined parameters of the journey factor missing indicator model (Model 3). This layout has been chosen to facilitate easy explanation of the DAG and for visual simplicity (as DAGs are notoriously known to look complicated). While implementing the model however, the parameters associated with each covariate are explicitly specified. A second unconventional aspect of the DAG presented in Figure 7.4 is its layout, which was structured to be visually easy to assimilate the relationships between the variables. Traditionally DAGs are laid out from left to right (Pearl et al., 2016), starting with the exploratory variables, and then any confounders. In DAGs the variables are typically laid out in the sequence that they are crystalized during the data generation process. The far right of a typical DAG would then be the outcome variable.

In the DAG created and shown in Figure 7.4, F_i ('Freq') represents the count of passengers, X_i are the observed covariates, and J_i are the journey factors (split into observed and missing components). M_i are the journey factor missing indicator, which is a fully observed binary indicator variable made up of components M_i^{mis} and M_i^{obs}. When a journey factor value is missing, M_i is '0', and this forms M_i^{mis} , otherwise M_i is '1' when the journey factor is observed, thereby forming M_i^{obs}. The θ_F, θ_J, and θ_D have already been introduced in the previous paragraph. The left and right sides of the DAG respectively are models for observed and missing data values. F_i is seen as the outcome with confounders (θ_F, X_i, J_i) pointing towards it. This models the fact that the count of passengers is related to combined mobility system parameters, the variables of mobility, and the journey factor. The journey factor J_i is appropriately created to be a proxy confounder and a mediator, in the sense that it causes both the outcome F_i and the exposure M_i^{obs}, while also acting as a mediator between θ_J and F_i. The terminology typically used in explaining the relationship between covariates and variables in DAGs are reported in the literature (Tennant, 2018).

The DAG underpins the solution mechanism implemented in the MCMC Bayesian imputation as DAGs are created in the belief that they specify the mechanism through which the data was created. DAGs visually portray the combined Bayesian models implemented, and specify the relationship and functionality of each variables (as exposures, proxy confounders, confounders, mediators, colliders, outcome etc.) in a causal[80] framework (Tennant, 2018). The DAG also serves a primary role in model testing and causal search (Pearl et al., 2016). If for instance the DAG in Figure 7.4 has been created as portraying the mechanism of mobility data generation, then we can test the model. The path $X_i \rightarrow J_i \rightarrow M_i^{obs}$ for example indicates that X_i and M_i^{obs} are independent given X_i (Jensen, 1996). As such if we regress M_i^{obs} on X_i and J_i, and it turns out that the parameter of the variable X_i is not zero, then the DAG model is wrong. In such a case the causal model specified is incorrect and we could review such a model. By so doing for the other covariates and relationships, we test and develop our DAG to create the model which identifies the cause and effect of various drivers of mobility.

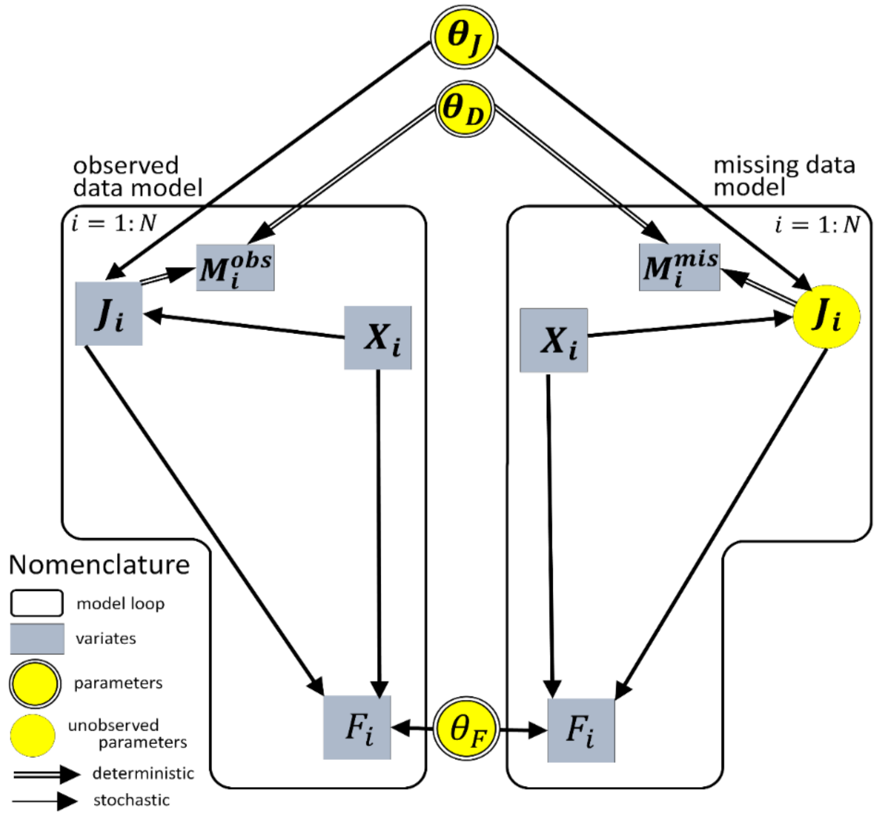


Figure 7.4 DAG main model including missing ‘Journey Factors’.

[80] The words confounders, mediators, outcomes etc. used to describe the functionality of the variables and covariates in a model are self-explanatory, with definitions detailed in the literature (Tennant, 2018)

7.4.3.2 Missing ticket validity period

The ticket validity period is the number of days during which a particular ticket product can be used. For example a single day return can only be used for 1 day, whilst a 7 day season ticket can be used for 7 days. This creates ticket validity periods of 1 and 7 respectively. As mentioned in Chapter 2 and Chapter 3, certain tickets are sold that are valid for say 70 days. On the ticket product these are described as being valid for between 60-90 days, so there would be no way of telling from the face-value of the ticket product that the ticket was a 70 day ticket. Under such conditions the ticket validity period would not be precisely known, and would be considered missing. These tickets products are described as flexible period tickets, and the actual period the ticket might have been used would be relevant information in establishing accurate passenger demand and mobility. As such, a method is developed for imputing the ticket validity period.

The flexible ticket products have codes which indicate the band of validity of the ticket. The MTA[81] are valid for between 30-60 days, MTL for 90-180 days, and MTH for 180-359 days, with no indication on the product description of the exact validity period of the ticket. For instance, two separate MTL tickets might actually have been used for 91 days and 171 days respectively, with no indication of these actual days on the ticket product. Passengers associated with these tickets would have contributed differently to the passenger demand on the network. For instance the 91 day ticket would be used for 3 ticketing periods[82] of 4 weeks each, whilst the 171 day ticket would have been used for over 5 ticketing periods. As such the 171 day ticket would be contributing to passenger demand 2 periods after the 91 day ticket would have expired. As the average daily passenger demand estimated on the network is the volume of passengers in a period (4 weeks) divided by the length of the period (28 days), the number of days of validity of flexible tickets would need to be ascertained to obtain objective estimates of passenger demand in specific ticketing periods. This section as such develops a Bayesian strategy for imputing ticket validity periods.

[81] The full meaning of these abbreviations (MTA, MTL, MTH, etc.) could not be established, but could have been historically adopted names with the actual meanings lost in evolution of the UK rail sector over time.

[82] This research is constrained to analysis in ticketing periods as the LENNON ticketing data supplied to the project are aggregated into ticketing periods of 4 weeks.

Conventional imputation strategies like ML, MI, and MICE would deduce the missing ticket validity periods by regressing the variables and covariates of the individual-level attribute rich ticketing data derived from m-IPF and GIS-GTFS. This would be in the assumption that the m-IPF and GIS-GTFS procedures yielded the individual-level data that are MAR, such that the available observed data are sufficient to explain the missing values. However, note that there is some additional information that would not be exploited by the ML, MI, and MICE procedures. These include the flexible periods of tickets like the MTA, MTH, MTL although not precisely known, are limited to within a time band. For example, we have information that MTA and MTH ticket products respectively have validity periods lying between 30-60 days, and 180-359 days. The advantage in using the Bayesian framework is that there exists simple constructs that enable us include such additional information in the regression exercise. In the BUGS Bayesian framework, many of such constructs are inbuilt, making it easier to build more comprehensive yet parsimonious models.

In the case of MTA tickets for instance which we know are limited to between 30-60 day of use, the BUGS Bayesian modelling framework includes this information by using the inbuilt censor construct $(C(v_{min}, v_{max}))$, which specifies that the exact value of the missing value is not known, but that they lie between the minimum (v_{min}) and maximum (v_{max}) validity periods which are additional information derived from the ticket product description. The exact value of a particular validity period is as such derived from a Bayesian regression model of the data, and by including additional information about the time limits of ticket use. In the Bayesian framework, these additional information can easily be incorporated thereby making for a robust model.

The Bayesian procedure for imputing the missing ticket validity period (for those ticket products with a non-fixed validity), is to account for the correlation between the other covariates (origin, destination, distance, cost, journey factor, etc.) using conditional regressions on the remaining not-yet-modelled incomplete covariates and any complete covariates. Table 7.2 is a formal model of the missing values in the ticket validity variable. The outcome and exposure variables are indicated using the convention that the first character (O or E) indicates if the variable is an outcome (O) or exposure (E). The second character indicates the type of likelihood function specified for the outcome, or the prior specified if the variable is an exposure. Any further constructs adopted in the model like truncation (T) or censoring (C) are indicated as the third character.

Table 7.2 Formal model of the missing 'ticket validity periods'.

Outcome explanatory (O) or Exposure predictor (E) variable	Model 4: passenger flows
Origin - station entry	E/N
Destination – station exit	E/N
Distance (journey, access, or egress)	E/N
Cost (monetary, or time)	E/N
Ticket validity period (days/months)	O/P/T
Group station flow	E/N

To accommodate the interests of Bayesian analysts, a pseudocode of the ticket validity period is included in Figure 7.5, illustrated as Model 4 within Figure 7.5. Model 4 in Figure 7.5 is added to the models in Figure 7.3 to enable imputation of the missing ticket validity periods. Note that the outcome variables used in previous models (Model 1 to Model 3 within Figure 7.3) are not included in this new regression in Model 4 as their influence on the missing covariates are already implicitly accounted for through the earlier models.

```
model {
## Model 4 - missing ticket validity (rates/offset) variable
    tickt.val[i] ~ dpois(mu.2[i])C($(v_{min}$ [i] ,$v_{max}$ [i] $)$
    log(mu.2[i]) <- etha.0 + etha.1[O.Code[i]] + etha.2[D.Code[i]] + ....   }
```

Figure 7.5 Missing Ticket Validity' script.

It is noteworthy to point out that additional model constructs are typically considered in the model development stages. For example, the Poisson regression model chosen to model the journey factors and ticket validity periods were considered for replacement using a likelihood function composed of a mixture of Normal distributions. The mixture of Normal distributions is premised on that passenger behaviour tends to be different when using fixed and flexible tickets. The mixture of Normal distributions captures this difference in passenger behaviour, whilst accepting that the passengers come from the wider pool of railway passengers. The change from Poisson to Normal mixtures (Kruschke, 2014, Plummer, 2003) yielded a nominal improvement in precision of the results, hence the Normal mixtures model was not adopted to generate the results presented.

The corresponding DAG for the missing ticket validity periods imputation model (Model 4) is shown in Figure 7.6. Observe that the earlier DAG of Figure 7.4 is similar[83] to that in Figure 7.7. The difference is the exclusion of the journey factor indicator model (which made limited impact on the overall model), and the inclusion of the observed and missing components of V_i (respectively on the left and right of the DAG). The inclusion of V_i is to reflect its conditional dependencies on the other covariates, forming the basis of the workings of the Bayesian imputation model. The θ_v is the parameter of the covariate for ticket validity period. The DAG specifies the conditional dependencies between the covariates $J_i, X_i, V_i,$ and F_i . These relationships can be verified from the covariates within the dataset to ascertain the validity of the DAG (as highlighted in Section 7.4.3.1). If the DAG does not reflect the conditional dependencies between the covariates and variables, then it is a wrong model of the data. In such a scenario, the DAG would need redesigning knowing where it went wrong.

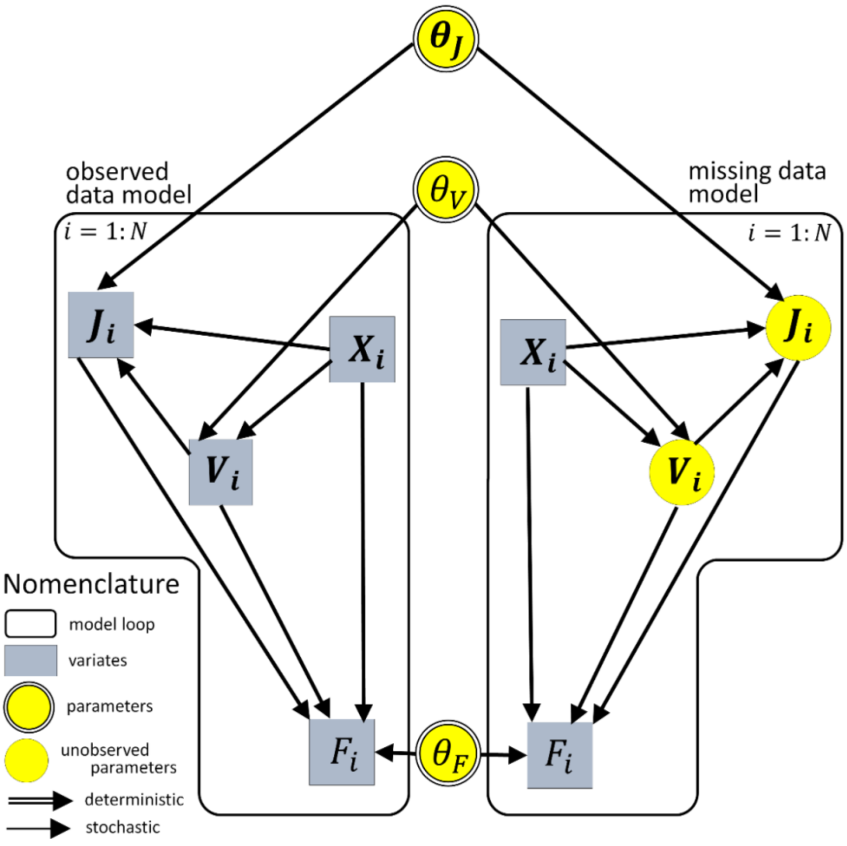


Figure 7.7 DAG main model including missing ‘Ticket Validity’.

[83] The layout of the DAGs are made similar to highlight the model development stages by increasingly adding relevant variables, while exploring the convergence efficiencies achieved for different model constructs.

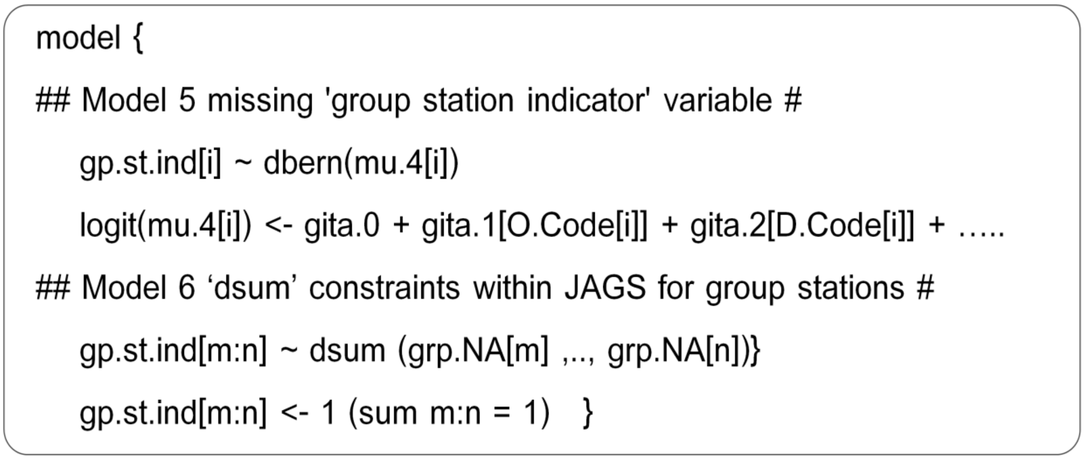
7.4.3.3 Missing group station flows

There are two stations in Wakefield called Wakefield Westgate and Wakefield Kirkgate. When many Network Rail tickets are sold for journeys to or from these two stations, the ticket product is ascribed Wakefield BR, and passengers can legitimately make journeys to either of these two stations. As a result, the two stations in Wakefield are described as group stations. As highlighted in Chapter 3, there are many group stations on the UK railway network. The group station tickets were primarily designed to make passenger journeys easier to plan logistically. Also, the original purpose for collating ticketing data within LENNON was for revenue appropriation to network operators, and not for passenger mobility analysis, making the need for precise information on passenger mobility less pertinent. The ticket products sold for flows to stations which are administratively clustered into a group are unique. For these tickets, a specific station name is not written on the ticket, instead the name of the group is written. In another example for instance, a ticket that is ascribed Leeds to Bradford BR implies that the ticket is valid for destination flows to Bradford Forster Square or Bradford Interchange. Forster Square and the Interchange although physically at different geographic locations, belong the same group of stations called Bradford BR.

When a passenger uses a ticket valid for entry to or exit from a group of stations, it is not readily possible on the face value of the ticket to know which one of the group of stations the passenger eventually used for entry or exit to the train network. Any one of the stations in the group would be a feasible choice. The issue that arises in using LENNON ticketing data for mobility analysis is that the precise flows to group station would need to be known in order to ascertain accurate passenger demand at each train station on the rail network. As a result, group station flows are ascribed as unknown and need to be imputed. The unique strategy developed in this research to identify the more probable station of entry and exit, is to augment the flows by assuming that the single flow to a group station now becomes multiple flows to all the stations within the group of stations. If the flow emanated from a group station and terminated in a group station, then augmented flows are created assuming the passenger travelled from all the stations at the entry end, and the same for all the stations at the exit end. A Bayesian imputation methodology is then developed as shown below, indicating that all the augmented flows are unknown, but that only one flow actually occurred so that the sum of the augmented flows to a group of stations is equals to one.

The Bayesian imputation model for flows to group stations is created as follows: having created the augmented flows originating and terminating at a group station, then, a binomial group station indicator variable is defined to be ‘1’ if a flow is precisely defined on the ticket product, and ‘NA’ otherwise. All the train station flows that are not associated with a group of stations are inherently precisely defined and would have an indicator of ‘1’, while the stations within a group would have indicators ‘NA’, requiring imputation. Two indicator variables are created one for the origin flows and another for the destination flows. The constraint that is imposed on the model is that the sum of the NA’s for a particular flow to or from a group of stations would sum to 1, with $NA \in [0,1]$. The Model 5 within Figure 7.8 is an implementation of such an imputation model for the group station flows. Model 5 is the Bernoulli regression model for the group station indicators with missing values. As the augmented flows to a particular group of stations would sum to ‘1’, to implement this and resolve which station attracted the flow, the ‘dsum’ construct within the JAGS software (Kruschke, 2014, Plummer, 2003) was used as shown in Model 6 within Figure 7.8. The ‘dsum’ ensures that within the Bernoulli model, the sum of the flows to stations within a group for a specific passenger OD flow results in a unit flow, as depicted by Model 6 within Figure 7.8. The full implementation of the ‘dsum’ construct has been included in Appendix D.6, and in the CD accompanying the thesis.

In the models within Figure 7.8 (in line with the convention established from the detail presented in the main model of Figure 7.3), the likelihood function is a Bernoulli distribution with parameter $mu.4$, and outcome $gp.st.ind$. The parameters of the Bernoulli regression are the $gita.i$, $(0 < i < N)$, where N is number of covariates in the regression model. Note the coefficients of the ‘dsum’ construct are the data values that are required to sum to 1.



```
model {
## Model 5 missing 'group station indicator' variable #
    gp.st.ind[i] ~ dbern(mu.4[i])
    logit(mu.4[i]) <- gita.0 + gita.1[O.Code[i]] + gita.2[D.Code[i]] + …..
## Model 6 ‘dsum’ constraints within JAGS for group stations #
    gp.st.ind[m:n] ~ dsum (grp.NA[m] ,.., grp.NA[n])}
    gp.st.ind[m:n] <- 1 (sum m:n = 1)   }
```

Figure 7.8 Missing ‘Group Station indicator’ script.

The DAG incorporating the missing data model for the group station flows is shown in Figure 7.9. The θ_G represents the joint parameters of the Bernoulli imputation model, with the group station indicators G_{pi} as the outcome variable. The DAGs in Figure 7.9 is similar to those in Figure 7.4 and Figure 7.7, with the difference that an imputation model has been included for the missing group station indicator variable. The DAG in Figure 7.9 illustrates the conditional dependencies between the covariates of the main regression model and the imputation models. The DAG was developed in part by assessing the relationships between the covariates to ensure that the DAG model adequately portrays the data relationships observed. The DAG in Figure 7.9 includes the main mobility model and all the relevant the models for imputation of the missing journey factors, the ticket validity periods, and the group station indicators. Having developed the DAG model by studying regression expressions represented by the conditional dependencies of the data variables and covariates, the DAG serves as a visual and analytical tool aiding in developing an understanding of the causal processes of data generation. This is a step further on from the traditional statistics inference paradigm of simply establishing how correlated sets of variables are, and identifying parameters that describe the joint distribution of these variables. The causal inference paradigm of developing a DAG answers questions about how changes in any one variable causes changes in another, and how much this causal change might be.

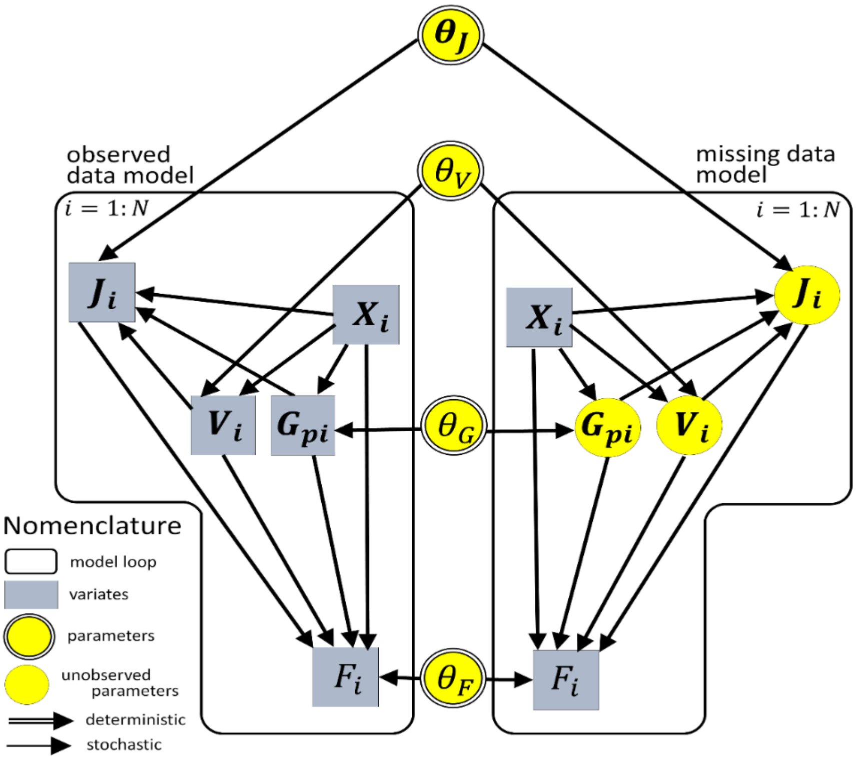


Figure 7.9 DAG main model including missing 'Group Station' indicators.

7.5 Result of Bayesian imputation

The full software scripts for implementation of the Bayesian imputation models are described in the Appendix D. The details of the software scripts are included in the CD accompanying the thesis. The scripts have been included to enable the results to be reproduced. The results produced are values for the parameters of the mobility model, and the missing data values for the journey factors, ticket validity periods, and indicators for group station flows. The imputation results are at individual-level (or group level) for each tuple with missing values. (Recall that the micro-level synthetic population was created from spatial microsimulation and GIS-GTFS network simulation). The Bayesian results enable the level of uncertainty in each of the results to be measured from the associated credible intervals. The results also present an assessment of the accuracy, efficiency, and representativeness of the Bayesian model. The individual-level imputed data enables an identification of the idiosyncratic behaviour of each individual within the railway network.

Each of the covariates with missing values have about 10% missing data, and this represents over 23,000 individual with missing attributes. The journey factor variable for instance has about 7.5% missing values, and this would imply over 19k individual results for the imputed journey factors. It would be impractical to present all the 19k results, but instead only a few typical and representative results are presented. The results can easily be generated by running the model scripts included in Appendix D.

The first set of results presented is for the distance decay parameter of the main model which is a negative Binomial model. The distance decay portrays the propensity of passengers to travel shorter distances on the network. In typical spatial interaction mobility models (Batty and Mackie, 1972, Besag, 1974, Dennett, 2012, Ewing, 1974, Wilson, 1971), the distance decay parameter is negative to reflect a passenger’s resistance to travelling longer distances. The second set of results is made up of three sub-sets. These are the imputation results for the journey factors, the ticket validity periods, and the group station indicators. The imputed journey factors are derived from a mixture of Normal model and results are presented for different ticket products. The imputed ticket validity periods are generated using a Poisson regression model, and similarly results are presented for different ticket products. The imputed group station indicators are generated using a Bernoulli model. Recall the indicators identify the passenger flow to a particular station within a group of stations.

7.5.1 Distance decay parameters

In order to identify the distance decay parameter, the count of passengers traversing from origin and destination Postcodes and accessing the train network at the various entry and exit train stations formed the outcome variable and yielded distance decays shown in Figure 7.10 and Figure 7.11. The pertinent covariates are made up of the origin (Postcode of usual residence) and final destination, station entry and exit points, travel purpose, travel distance, journey factor, ticket validity period, group station indicator. An additional range of socio-demographic variables were included in the imputation pertaining to each passenger, these are the travel purpose, income, age, gender, household type, number of cars in household, and children, and ethnicity. Additional covariates included number of stops, travel time, arrival and departure times to the rail network, station access and egress distances. The list of all attributes considered in the main and imputation models are reproduced in Table 7.3, and in Appendix D.7. These attributes are derived from over 160 variables inherited from the Census, NRTS, LENNON, and GIS-GTFS model. In identifying the distance decay, two modelling configurations were investigated for the journey factors. In the firs, the outcome variable in the missing data model for the journey factors was assumed to be a Poisson distribution (see Model 2, Figure 7.3). In the second, the outcome variable was assumed to be a mixture of Normal distributions. Typical results are shown in Figure 7.10 and Figure 7.11.

Recall that the essence of Bayesian modelling is to simulate a Markov chain whose stationary distribution is the sought target distribution. Usually the target distribution is the posterior distribution of the model parameters being estimated. Before using the simulated chains from the Bayesian MCMC to obtain estimates of parameter values, there are affirmations that need to be made to ensure that the results are objective. The simulated chain has to converge to a stationary distribution for objective deductions to be made about the parameter values. This convergence can be visually ascertained using the trace plot, typical of the top left plots in Figure 7.10 and Figure 7.11. The trace plot (where the vertical axis is the parameter value) shows the history of a parameter value as the iterations of the chain evolves. In the trace plot, the parameter values are plotted as they evolve (with the evolution of the chain along the horizontal axis). Typically a number of chains are simulated simultaneously to facilitate a more efficient investigation of the parameter space. If a chain becomes stationary, then majority of the parameter values searched will be scattered around the mean parameter value, with no systematic long term trends, and the chain would have converged. Typically when a chain has converged (after an initial unsettled period) the trace would resemble a so called caterpillar. If multiple chains are simulated and have converged, then there would be no systematic difference in their scatter. The top left plot in Figure 7.10 resembles a trace plot that has converged enabling us to have more confidence in subsequent deductions from this trace. The three trace plots are scattered on top of each other. Under such conditions, the parameter values deduced from a density distribution of such a chain (as displayed in the bottom right plot of Figure 7.10) would statistically represent a correct estimate of the parameter distribution.

The trace plot can visually reveal a range of issues with the Bayesian MCMC procedure. If for instance the step size of the random walk iterations is too small so that it takes many iterations for the chain to traverse across the mean posterior distribution, then the chain would require a very long time to explore the entire parameter space and visually the trace plot would exhibit long term trends reflective of parts of the parameter space that are investigated. If parameter deductions are made from such a trace plot, the mean parameter value would be subject to the part of the plot that was used to create the parameter density plot. A remedy would typically then be to run the chains for a much longer period, and the trace plot would have revealed this requirement. The trace plot in the top left of Figure 7.11 exhibits such non-convergence. It is seen that the three chains do not overlap indicative of non-convergence. Under such conditions the chains would be required to run for much longer before they converge, and this is indicative of an inefficient MCMC. Under these conditions, the analyst would be precautious in subsequent deductions made from such trace plots. The parameter values shown in the bottom right plot of Figure 7.11 hovers around a range -2.0 to 0, with a pronounced amount of uncertainty in the actual peaks. However, such parameter density distributions could still be useful in revealing the nature of the mixture of Normal distributions that define the mobility mechanism.

In summary, the Poisson model adopted for the journey factors yields trace plots that mix well as shown in the upper-left plot in Figure 7.10. The trace plots shown in the upper-left plot of Figure 7.11 show that the mixture of Normal model when adopted for the journey factor variable yields less efficient chains that are not stable for the same number of iterations. The Poisson model is a more efficient model than a mixture of Normal when imputing the journey factors.

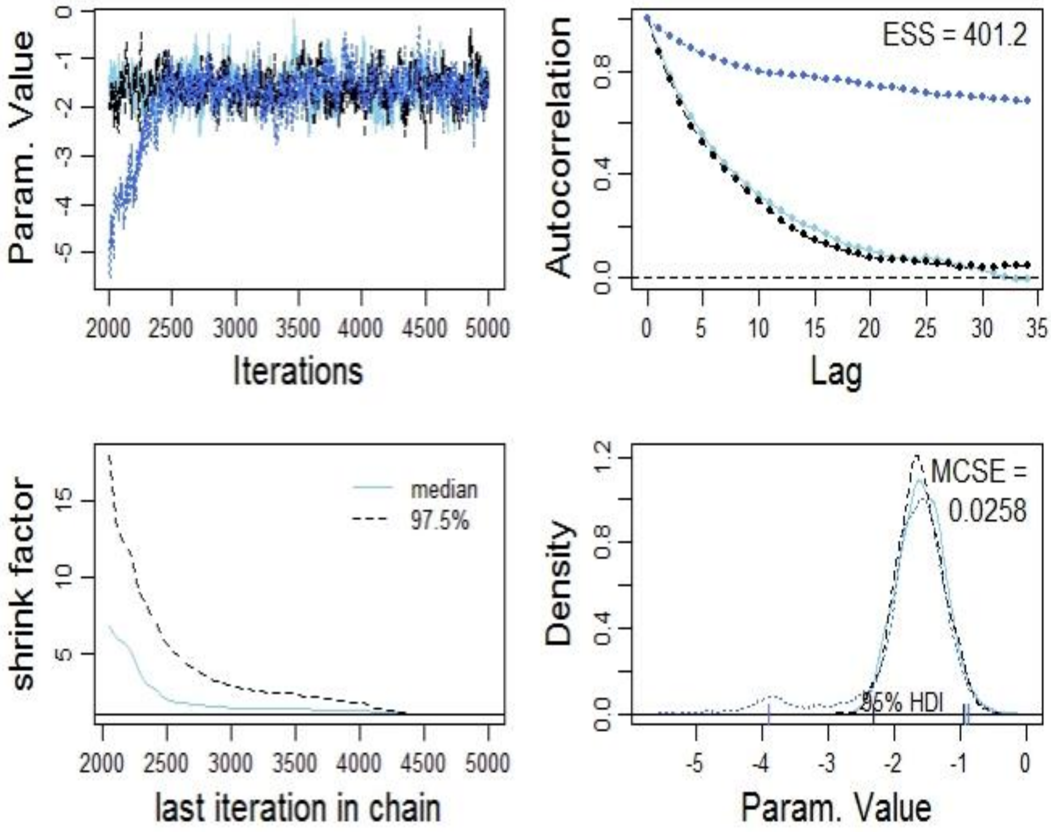


Figure 7.10 Distance decay – Poisson 'Journey Factor'.

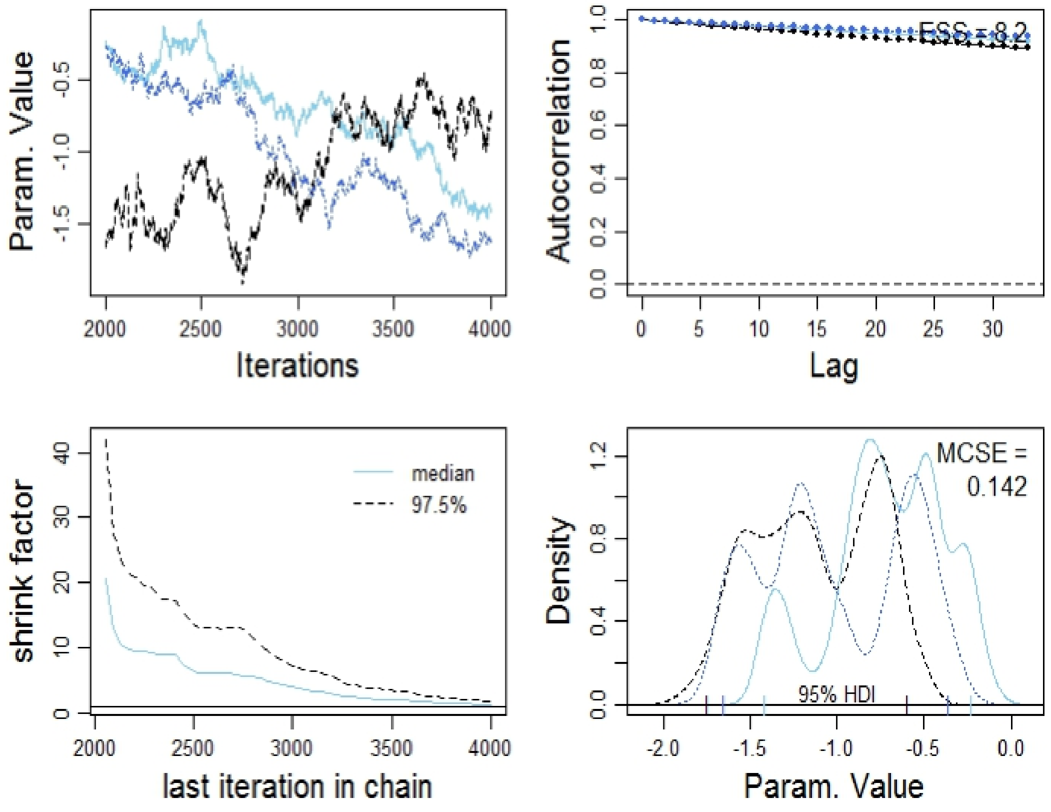


Figure 7.11 Distance decay – mixture of Normal 'Journey Factor'.

A property of Markov chains (MC) is that during iteration, a sample on the chain only depends on the state attained in the previous step of the chain, and on none other. This is a unique attribute of the random-walk[84]-like MC that ensures an efficient exploration or sampling of the parameter space to attain an equilibrium distribution of a targeted distribution. When a Bayesian MCMC chain has converged and is stable, the autocorrelation from one chain sample to another would indicate dependence of any one sample on a subsequent sample, but this autocorrelation would decrease rapidly (from an autocorrelation value of 1) as the distance (lag) between the chain samples increase. This is a required property of an efficient Markov chain. The top right plots in Figure 7.10 and Figure 7.11 are plots of the autocorrelation against different lags. For example the top right plot in Figure 7.10 where a Poisson model is used to impute the journey factors, the autocorrelation of two out of the three chains is high when the lag is low, and the autocorrelation rapidly drops with an increase in lag. One of the autocorrelation plots in Figure 7.10 however, does not drop rapidly indicative of an issue with that particular chain. An investigation of the corresponding trace plot reveals that earlier parts of that chain between iterations 2000-2500 had not converged. An inclusion of this part thereby yields an autocorrelation plot that does not show strong independence of subsequent chain samples as the iteration progresses. Autocorrelation is important as it is indicative of how much information is available in the Markov chains. For example, sampling N iterations from a Markov chain that has a high autocorrelation yields less information about the stationary distribution than would have been obtained from a sample of N independent draws from the stationary distribution. As such, the autocorrelation plot is indicative of the efficiency of the Markov chain. The intuition behind adopting autocorrelation as an assessment of efficiency of MCMC chains is simple: if a statistically representative sample is desired from a population it makes more sense taking independent samples from the population. A sample taken of members of the same family within the population would not be representative of the wider population. As such a sample would be auto correlated coming from individuals related by being in the same family. An observation of the autocorrelation plot on the top right of Figure 7.11 shows that the chains are highly correlated even when the lag is as high as 30. This

[84] The random walk (RW) is distinct from the Markov chains (MC) in that whilst MC's iterations are only dependent on the previous chain, the RW iterations are dependent on no other chains.

is indicative of a poor and non-independent chain of samples. As such, inference on parameter values deduced from such a chain (and shown in Figure 7.11) would only be indicative, and a poor reflection of the true parameter distribution. As such, inference about the quality of the Bayesian models can be made by visually inspecting the autocorrelation plots.

An additional indicator that is derived from the autocorrelation plot is the effective sample size (ESS). Recall that the essence of the MCMC is to sample the parameter space efficiently. The MCMC iterations are essentially made up of an exploratory stage (the Monte Carlo) and an optimisation stage (the Markov chains). If the MCMC is efficient, then the majority of the sample candidates would be accepted, and this reflective of the quality of the optimising Markov chain. A high candidate acceptance rate implies that the effective independent samples taken amount to a lot. In the ultimate case where there are no rejected samples, the effective samples would be equivalent to the number of iterations. A measure of the efficiency of the chains is as such the effective sample size, called the ESS[85] (Neal, 2001). This additional information is revealed within the autocorrelation plot. For example, embedded in the autocorrelation plot of Figure 7.10, is the ESS which is 401, indicating that out of 5000 iterations, 401 effective samples where generated. In Figure 7.11 however, the ESS is low at 82, indicative that the corresponding chain would have to run five times as long as the chains derived from the Poisson model used to generate Figure 7.10. The ESS is a measure of the number of independent samples that would need to be taken to achieve the same inference as would be achieved from the trace plot. Alternatively, the ESS can be interpreted as the length of chain you would have left over if the chains are thinned (coarsely sampled) to remove the autocorrelation.

In summary, the Poisson model which has a higher ESS (401.2) converges faster than the mixture of Normal model which has a lower ESS (8.2), indicative of the requirement for long run times for convergence to be achieved in the mixture of Normal model. Referring back to the trace plots, the mixture of Normal model requires longer run times to stabilize and for any inference to be deemed statistically representative.

[85] The ESS is typically defined as the number of independent samples that would need to be taken to achieve the same presented in the trace plot. Alternatively the ESS can be interpreted as the length of chain you would have left over if the chains are thinned to remove the autocorrelation.

To further improve the confidence in the results of the MCMC Bayesian models, Gelman-plots are created as shown in the bottom left of Figure 7.10 and Figure 7.11. These plots are dependent on multiple chains[86] being run during the MCMC, and develop a strategy for comparing these chains as a basis for improving the confidence in the parameter estimates. When multiple chains are applied, and these converge to the same location, a confirmation of parameter stationarity can be infered by conducting the Gelman-Rubin diagnostic (Kruschke, 2014, Plummer et al., 2006). This diagnostic computes the within chain variability and compares this to the across (between) chain variability. The intuition is that if a stationary solution is reached, the variability across the chains should be small, and comparable to the within chain variability. The Gelman-Rubin statistic computes the variance of all the chains combined and compares this with the mean of the variance within each chain. Values that is substantially higher than 1 are indicative of non-stationarity and as such non-convergence. Typically, the parameter distribution of non-converged chains tends to be broad, with the potential to shrink if the MCMC simulation is run for longer. As such, chains with high Gelman-Rubin statistic are termed to have high shrinkage factors. The Gelman-plot reveals how the shrinkage factor changes as the number of iterations increases. For instance, the plots on the bottom left of Figure 7.10 and Figure 7.11 are these Gelman-plots. The plots show that whilst the model with the Poisson journey factor attains a shrinkage factor of 1 nearer 4300 iterations (as seen in Figure 7.10), the model with the mixture of Normal attains a shrinkage of 1 at 4000 iterations (as seen in Figure 7.11). The important contribution of such Gelman-plots is that they indicate that after about 4000 iterations, the within-chain and between-chain variability are comparable implying stationarity in the chains. This is indicative that the burn-in period of the chains should be set at over 4300 iterations in order to expect stationary and converged chains, thereby ascribing confidence in the parameter results.

[86] The chains are typically designed to start from different initial values (also enabling an assessment of the sensitivity of the model to initial starting points). Models that are long-run sensitive to initial values cannot be considered to be stable. When the MCMC iteration is started and the chains are created, it many a time takes a while for the parameter space to be explored to traverse from the initial value to a stationary parameter solution. It is common practise to discard the unstable region of the trace between initial value and the beginning of a stationary parameter trace. This unstable region which is discarded is called the burn-in.

A further visual assessment tool for Bayesian modelling is the density plots shown in the bottom right of Figure 7.10 and Figure 7.11. These plots are the smoothed histograms of the parameter values sampled from the trace plots (on the top left in Figure 7.10 and Figure 7.11). When a Poisson model is used for the journey factors, the density plot in Figure 7.10 shows that the three trace plots produce overlapping density plots, indicative of convergence of the chains within 5000 iterations. On the other hand, when a mixture of Normal model is used for the journey factors, the density plot in Figure 7.11 shows that the three plots do not overlap, giving a visual indication that the chains have not converged.

Further, the density plots display additional statistics for assessing the MCMC results, and these are the highest density interval (HDI[87]) and the numerical Monte Carlo Standard Error (MCSE). The HDI is the span of values that cover 95% of the parameter density plot. The parameter values within the band of the HDI are considered credible. The MCSE is indicative of the amount of scatter in parameter estimates. The HDI limits (shown by the vertical markers) for the three chains in Figure 7.10 (for the case of a Poisson journey factor model) are similar but have a level of variability due to the finite length of the chains. The HDI limits for the density plots in Figure 7.11 (for the case of a mixture of Normal journey factor model) show a higher level of variability, indicative that the chains have not fully converged.

The 95% HDI for the Poisson journey factor model on average is in the range $(-0.9 < 95\%\ HDI < -2.4)$, whilst that for the mixture of Normal journey factor model is nominally narrower at $(-0.4 < 95\%\ HDI < -1.7)$, indicative that despite the poorer chains in Figure 7.11, the results are indicative of those that would have been obtained had the chains been run for much longer (Kruschke, 2014). In many practical situations, when a model is complicated by virtue of the number of covariates and their corresponding likelihood function specifications, and when the various Bayesian constructs are deployed, the models could run for a substantial period before convergence. It is further not simple to establish the correct initial conditions for the variables, making it difficult to ascertain the length in time of the burn-in periods. For instance, the models used to produce the

[87] The HDI in Bayesian modelling is the span of values that cover 95% of the parameter density plot. The parameter values within the band of the HDI are considered credible. The 95%HDI is equivalent to the credible interval, to distinguish from the confidence interval used in frequentist statistics.

results presented in Figure 7.10 and Figure 7.11 ran for 15 days on the University of Leeds of Arc 3, a 2-Linux CentOS(7)-based supercomputer node with 24 cores and 768GB of memory, with 350TB usable storage (University-of-Leeds, 2017).

The MCSE is equivalent to the standard deviation of the chain divided by the effective sample size (ESS) which was introduced earlier. The MCSE is a measure of the convergence of the chains, as the scatter typical of non-converged chains would reflect in a higher standard deviation. This amount is then normalised by the ESS which is an indication of the size of sample that would have resulted had the modelling been performed using a frequentist approach. As seen from Figure 7.10 and Figure 7.11, using the Poisson model for the journey factors yield a lower MSCE (i.e. 0.026) than when a mixture of Normal distribution is used to model the journey factor (yielding an MCSE of 0.142), indicative of less variation and more stationarity in the former model.

Expectedly both models yielded negative distance decay parameters, as seen from the values of the horizontal axis of the parameter density distributions. The negative distance decay is reflective of the propensity of majority of the passengers to travel shorter distances to fulfil activities. Both distance decay credible intervals[88] (CI) lie in the range $-0.5 < CI < -2.4$.

In summary, the goal in an MCMC Bayesian modelling exercise is to obtain a posterior estimate of the parameter values. Considerations made in assessing the model include a measure of how representative the MCMC chains are of the posterior estimate being identified. Another consideration is the size of the chain to ensure that enough effective samples were taken to ensure that the results are accurate by adequately exploring the range of potential results, and that these are reasonably stable to reflect convergence. Another consideration is the efficiency with which the results were attained, as this is reflective of the quality of the model, and ensures the results can be reproduced in a limited time frame.

[88] Note that in the Bayesian context, the credible interval[88] can be loosely thought of as the range of parameter values that represents 95% of the most credible results. Note also that the high density interval (HDI) in Bayesian modelling is subtly distinct from the CI, and represents the span of values that cover 95% of the parameter density plot.

7.5.2 Imputed covariates

The imputation results are presented in this section for the journey factors, ticket validity periods, and flows to group stations. For each set of results, the trace plots are presented, as well as the autocorrelation, Gelman shrinkage factor plots, and parameter density plots. These plots have been discussed exhaustively in Section 7.5.1, and so a repeat of such detail is not presented in this section, instead reference is hereby made to Section 7.5.1. For each set of results presented, an assessment is made of the level of scatter about the mean in the trace plots to infer stationarity of the MCMC chains. This scatter in the trace plots is sometimes referred to as mixing. The autocorrelation plots are created to enable an assessment of the quality of the MCMC chains, ensuring that each parameter sample is effective. Recall the property of Markov chains (M-C) that a sample on the chain only depends on the state attained in the previous step of the chain, and on none other. When such independence is not achieved, the effective sample size is low, then necessitating larger sample sizes to make objective inference. The Gelman-plots created reveal the shrinkage factor of the chains which give indications about the number of iterations required to achieve convergence of solution. In such a state the parameter density plot would have narrower (more precise) peaks and such peaks would be said to have shrunk in breadth. Finally, the estimated parameter density plots for the missing values are assessed to reveal the relevant credible intervals of solution. In each case, just like in the case of the main mobility model presented in detail in Section 7.5.1, three sampling chains are simultaneously run.

7.5.2.1 Missing journey factors

Recall that the imputation model for the journey factors is presented as Model 2 within Figure 7.3. It is a Poisson regression model with truncated parameter values. The outcome variable is the journey factor which is being imputed. While developing the Bayesian imputation model, recall that the journey factor imputation model excludes the outcome variable of the previously developed main mobility model (i.e. the passenger count variable). There are about 19,000 missing data values in the journey factor variable, so only a sample of typical results are presented along with the software code (in the Appendices) used to implement the model. These detailed specifications of the journey factor imputation model, as well as the priors and the initial conditions of the variables are discussed within Appendix D.2 to Appendix D.5. The typical Bayesian results for the imputed journey factors are shown in Figure 7.12 and Figure 7.13.

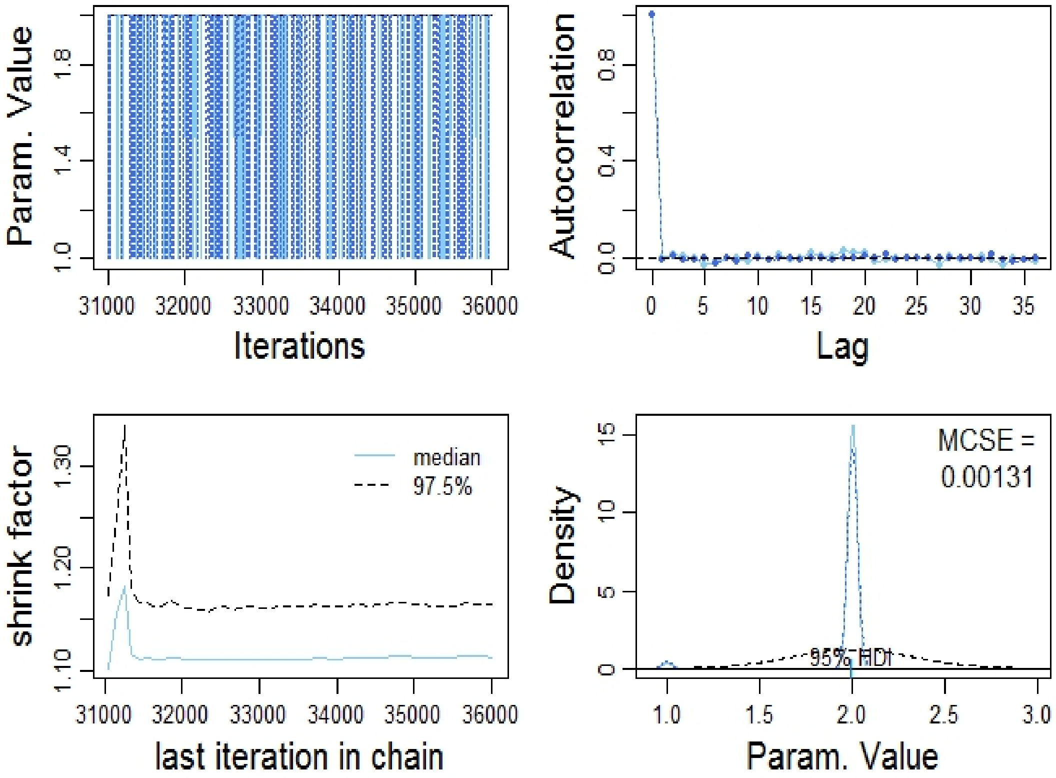


Figure 7.12 Journey factor for a 7 Day Season ticket (typical result).

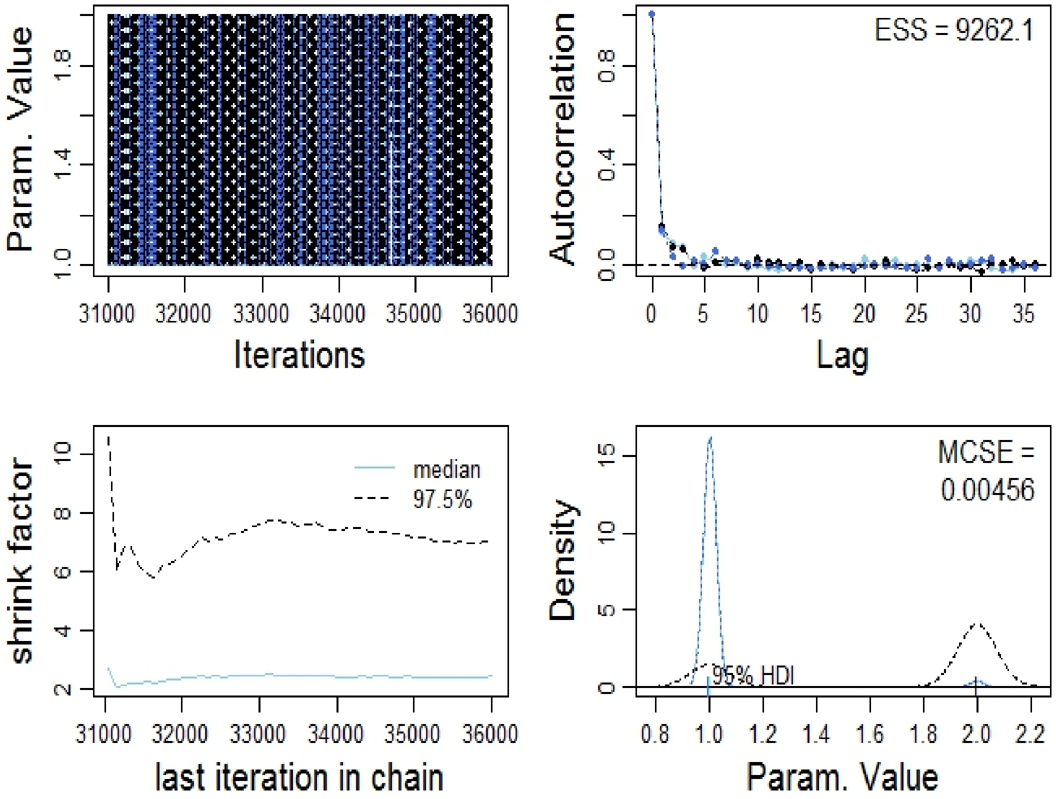


Figure 7.13 Journey factor for a Flexible Season ticket (typical result).

The results presented in Figure 7.12 and Figure 7.13 are journey factors for two different ticket types. Figure 7.12 is for a 7 day Season ticket, while Figure 7.13 is for a flexible Season ticket. As seen from the overlapping trace plots in the top left graphs, there is a scatter about the mean, this is indicative of adequate mixing of the chains, in turn indicating that the chains have converged to a stationary solution. The stationary scatter plots indicate that the chains are representative of the posterior distribution, and that there are no pronounced influences of the initial values. The autocorrelation plots show rapid drops as the lag between samples from a chain is increased. This is indicative of the independence of samples, as expected from a well behaved Markov chain process. As a result of this efficient MCMC sampling, the effective sample size (ESS) for both plots as numerically high (at 12785.1 and 9262.1) as indicated in the top right plots of Figure 7.12 and Figure 7.13. The large enough samples (due to high ESS) ensure that stable and accurate estimates are made from the distributions of the chains. The Gelman-plots in the bottom left of Figure 7.12 and Figure 7.13 indicate that the within chain variance becomes comparable with the across chain variance yielding values between 1.1 and 2.0 as seen from for the respective Gelman-Rubin statistics (shrinkage factor) at about the 31500 iteration mark. This indicates that the chains become stationary and converge after about the 31500 iterations.

The density plots presented for the missing journey factors in the bottom right plots in Figure 7.12 and Figure 7.13 respectively, are sharp, overlapping and show normal distributions. This indicates that the results can be considered stable, accurate, and representative of the posterior target. The density plots of the parameters do not traverse the zero and are as such objective. The median (and modal) values presented in Figure 7.12 and Figure 7.13 are respectively 2.0 and 1.0. The deduction is that the railway passenger using the particular 7-day Season ticket has a propensity to use it 2 times a day, whilst the passenger that used the particular monthly ticket used it on average once daily. Note the small peak at 2.0 in Figure 7.12 perhaps a residual effect of separating flexible return tickets into outward and inward journeys. It is noteworthy to point out that the imputed parameter values are of the order of magnitude of the journey factors adopted for such ticket products in the rail industry (SDG, 2012, SDG, 2016). The Bayesian journey factors add detail by deducing values for each individual ticket, thereby capturing the heterogeneity in individual mobility behaviour. The precision obtained at individual-level is an improvement on the aggregate proxy values hitherto used in the rail industry (Taylor, 2013b).

7.5.2.2 Missing ticket validity periods

Typical results for the ticket validity periods are shown in Figures 7.14 and Figure 7.15. In Figure 7.14, the ticket validity period estimated is 185 days for an MTH ticket described as a standard Season ticket valid for between 181-359 days. In Figure 7.15 on the other hand, the ticket validity period estimated is 32 days for an MTA described as a Season ticket valid for between 30-60 days. Current industry infilling ascribes values of 180 days and 30 days respectively to these tickets. The credible interval (95%HDI) deduced from Bayesian imputation of the MTH and MTA tickets are respectively $(180 < 95\%HDI < 212)$ and $(< 95\%HDI < 42)$. These credible intervals encompass the 180 days and 30 days adopted in the railway industry. The advantage of the Bayesian imputation over the ad-hoc rail sector infilling is that the missing values are deduced by taking cognisance of the data generation process thereby better reflecting the mobility mechanism in the infilling process. The Bayesian imputation also has an advantage of giving a measure of the uncertainty in the imputation.

A generic investigation of the imputation results shown in Figure 7.14 and Figure 7.15, shows the top left trace plots exhibiting varied levels of mixing. The results for the MTH ticket (shown in Figure 7.14) indicates a trace plot scattered about a mean value but with visual signs of truncation at the lower extreme of the chain where the parameter values are lowest. The MTH chains also exhibit a level of wandering in one of the chains. The MTA ticket on the other hand (shown in Figure 7.15) reveals a scatter and mixing in the trace plots with no obvious signs of truncation. Perhaps the MTH ticket seems to exhibit heightened influence of the truncation as the lower truncation limit is close to the missing value (within 3%). The MTA ticket on the other hand had the lower truncation limit removed to enable enough samples to be generated. The stationary nature of the trace plot for the MTA ticket is indicative that the MTA ticket chains have reached a stationary convergence stage, albeit not including any lower truncation limit.

The autocorrelation plots show that for the MTH ticket 2 out of 3 chains exhibited a depreciation in autocorrelation with lag. The non-converging chain could be due to a poor choice of initial condition, as this is also reflected in the corresponding trace plot that occasionally drifts to high parameter values. The MTH tickets have a low ESS (89.2) which is indicative of the low quality of the MCMC chains. The MTA ticket on the other hand shows a rapid drop in autocorrelation with lag, and these results in higher ESS (829) which is indicative of high quality MCMC chains. The Gelman-plots which are the bottom left plots in Figure 7.14 and Figure 7.15 show that the shrinkage factors of MTH ticket reaches the 1.0 at the 4000 mark, indicative that the threshold of the burn-in should be about 4000 iterations. This is indicative that the chains for the MTH ticket have not completely converged to a stationary solution. The shrinkage factor of the chains for the MTA tickets on the other hand reaches a value of 1.0 at about 2500 iterations. This is indicative that the burn-in period is lower than for the MTH ticket, and that the chains do converge within the iterations conducted (i.e. <5000 iterations).

The density plot for the imputed MTH ticket showed in the bottom right plot in Figure 7.14 exhibits density distributions that do not overlap precisely. This is symptomatic of an MCMC solution that has not converged. There is a skew in the density toward the 180 mark, perhaps a result of the lower truncation limit applied to the imputation model of the MTH ticket validity period. The variability in the density plots and the low effective sample size results in a high MCSE value at 1.29. The imputed MTA ticket on the other hand has a lower MCSE (at 0.204), and this is indicative of density plots with low variability and chains with high ESS values. The indication from the density plots is that the application of the lower truncation limit on the model for the imputed ticket validity periods, results in MCMC chains that require much longer to run. Such longer running MCMC chains can potentially yield more precise and accurate results as seen by the narrower density plot for the MTH tickets. In many practical modelling scenarios, it would be a trade-off between the available time to fully develop and run the model, and the requirement for an indicative solution to the particular problem.

In summary, the validity periods imputed for flexible tickets as shown in Figure 7.14 and Figure 7.15 are more reflective of passenger behaviour than proxy aggregate values assumed for behaviour of passengers in current rail sector applications. The ticket validity periods derived are influenced by the wider ticket dataset, inherently better reflecting the wide range of passenger behaviours. These imputed ticket validity results better reflect the actual dynamics of ticket use as each passenger’s behaviour is conditionally independent of others, and reflective of the particular passenger circumstance, and as such are typically variable. The current rail industry standard would be to assume that tickets of validity between 30-60 days are used for 30 days, and a similar proxy factor used to decide for all other flexible tickets. These are not reflective of the heterogeneity and idiosyncrasies associated with individual passenger ticket use behaviour.

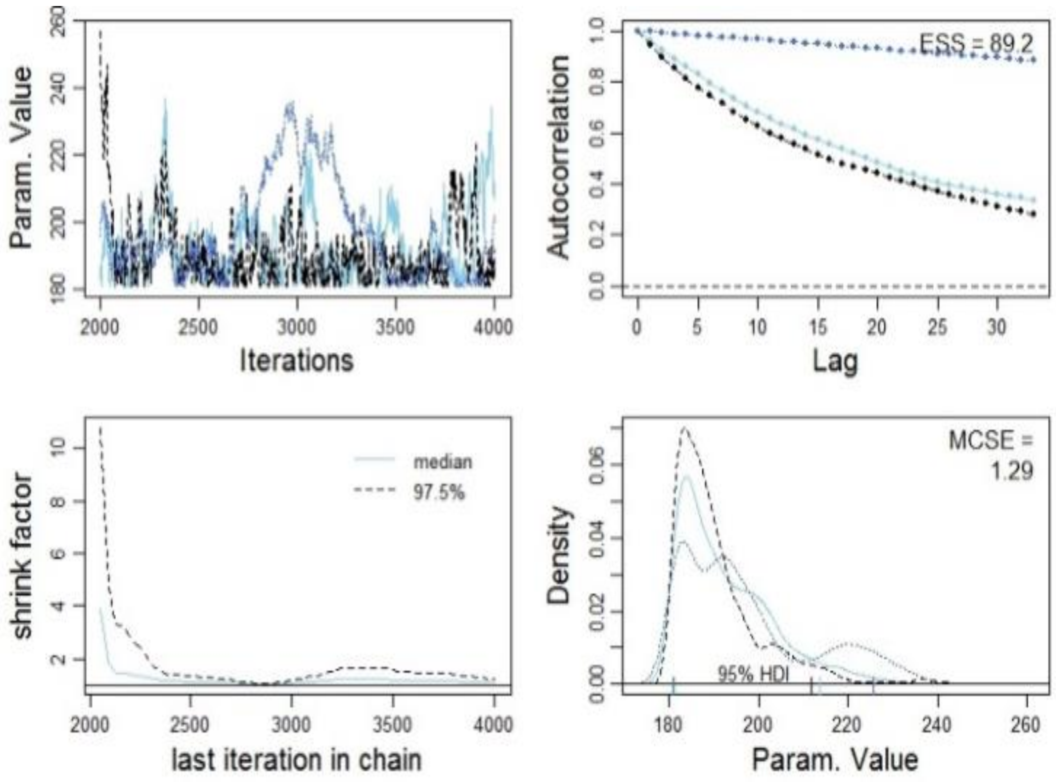


Figure 7.14 Ticket Validity MTH ticket (typical result).

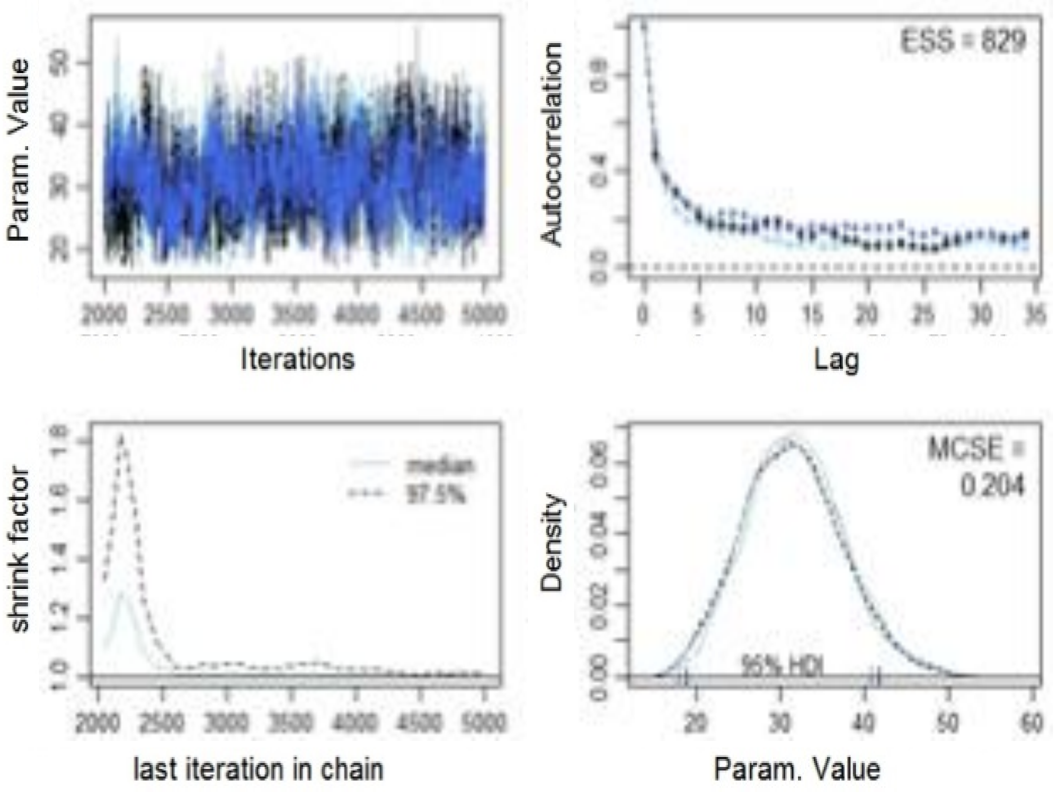


Figure 7.15 Ticket Validity MTA ticket (another typical result).

7.5.2.3 Missing group station indicator

Typical imputation results for the group station indicators are shown in Figure 7.16 and Figure 7.17. Recall that if a single flow occurred to a group station, then augmented flows are created to each of the stations in the group, in the proviso that the sum of these binary group station indicators will be 1. In essence, if there are 3 stations in a group and a single flow occurred to these, then we would create three flows, one to each station. We would also create group station indicators which are unknown (NA). The Bayesian imputation will then impose a constraint such that the sum of the NA's will equal 1, whilst the NA's are the outcome variable with a Bernoulli (or Binomial) likelihood function. The Bayesian model for the missing group station indicators is depicted in Model 5 and Model 6 within Figure 7.8. The detailed software code is described in Appendix D.2 to Appendix D.5.

Typical results shown in Figure 7.16 and Figure 7.17 indicate mixing in the chains as seen in the upper-left plots. The autocorrelation values also show marked drops indicative of convergence of the solutions. In some instances as seen in Figure 7.17, some of the chains exhibit a non-rapid drop in autocorrelation with lag. These are perhaps as a result of sensitivity of the procedure to initial values. The ESS values are typically high at about 10% of the number of iterations. The Gelman-plots exhibit shrinkage factors that drop to values close to 1.0 following about 2500-3000 iterations. This is indicative that the results would improve further had the burn-in period been extended to about 3000 iterations.

The density plots exhibit low MCSE (typically at about 0.03) indicative of chains that converged to stationary solutions. Many of the results show a split in the density plots between the binary values of ‘0’ and ‘1’. The results could be interpreted as ‘1’ or ‘0’, dependent on the value at which the highest density occurred. For instance, the results shown in the bottom right of Figure 7.16 and Figure 7.17 are density plots for flows to group stations Kirkgate and Westgate. In Figure 7.16 for Westgate station it is observed that the density plot has the higher peak at ‘1’, indicative that the flow occurred to Westgate station. In Figure 7.17 for Kirkgate station the higher peak also occurred at ‘1’ indicative that this imputed group station flow was to Kirkgate station. An alternative interpretation would be to leave the densities as they are, and adopt the median density values as the proportion of flows to a particular station. The sum of such flows is seen to still sum to ‘1’. Using such an interpretation, the cumulative sum of flows to a particular station would be indicative of the passenger demand to that station.

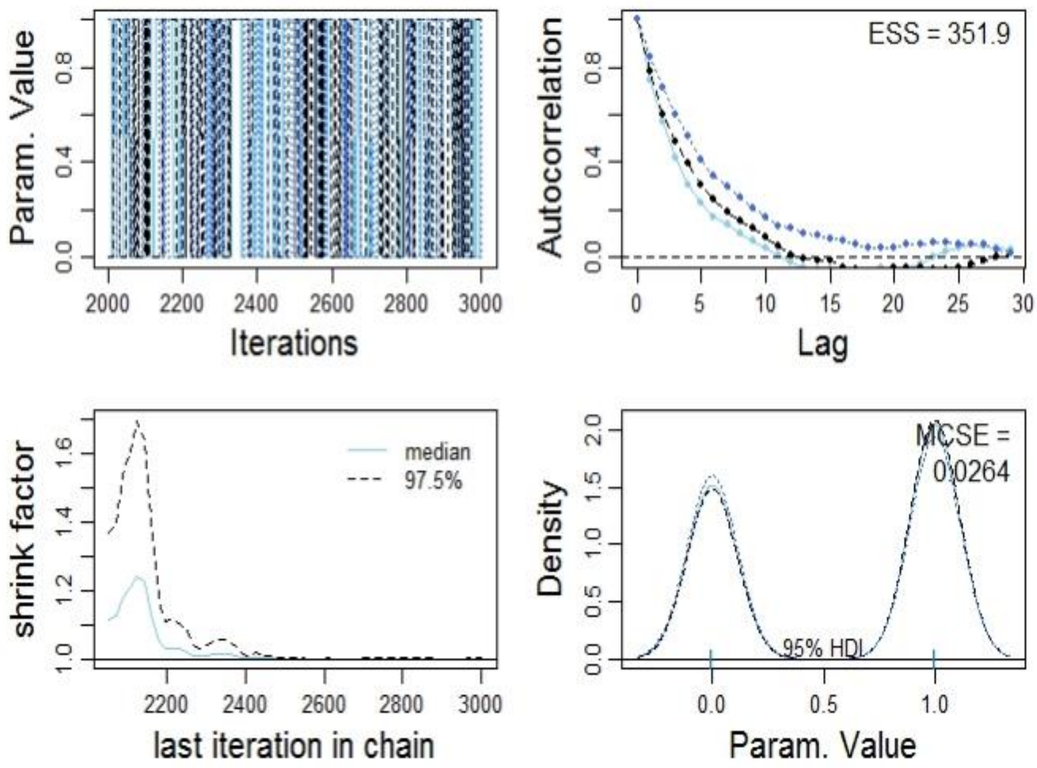


Figure 7.16 Group Station indicator (typical example).

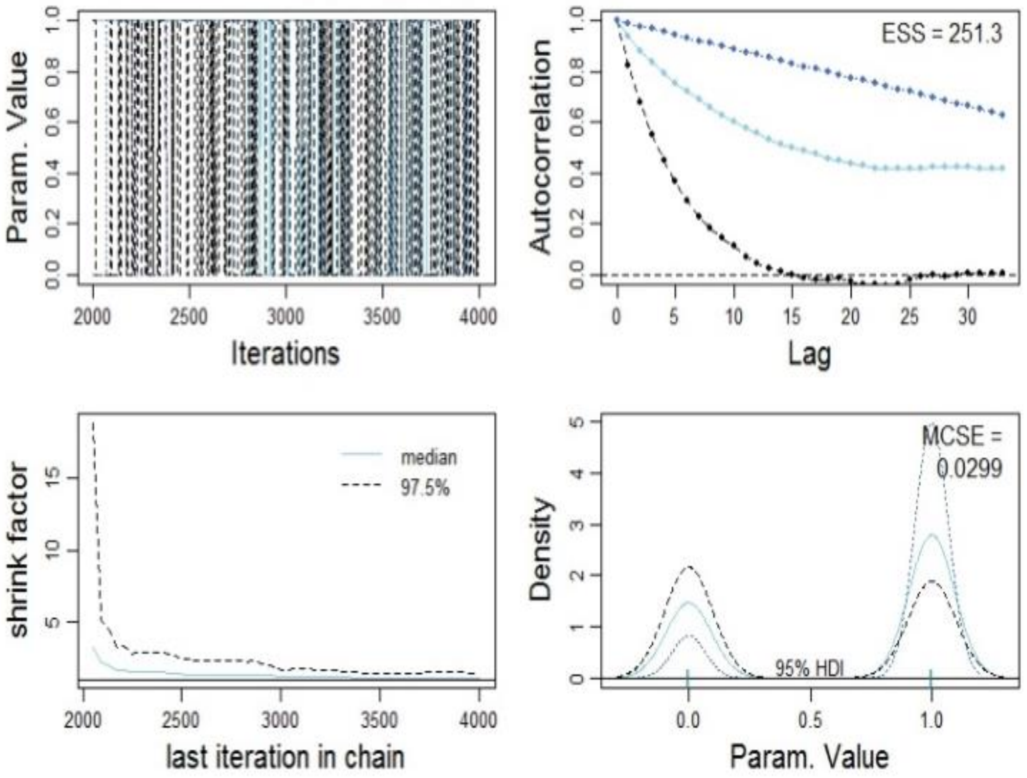


Figure 7.17 Group Station indicator (another typical example).

7.6 Discussion and remarks

During Bayesian imputation of missing values, the strategy has been to include as many relevant covariates yielded from spatial microsimulation and the GIS-GTFS network model. It can be conceived that Day cards or Return tickets serve a particular purpose and when compared to other ticket products (like Season tickets), would have journey factors reflective of the ticket product description (as demonstrated by the Little's test in Chapter 3, Section 3.2.3). The choice of a particular ticket product is reflective of the nature of mobility activity the passenger is desirous of fulfilling. This implies that for instance, the missing journey factors for season tickets can be inferred from knowledge of ticket product, activity being fulfilled, and type of passenger. The indication then is that the further inclusion of say the ticket price attribute in the imputation would enable a feedback of the influence of prize on the propensity for higher journey factors. In essence, a fuller spectrum of exogenous and endogenous attributes would be more objective in imputation of missing data values as postulated by Rubin (Rubin, 1976). This is the concept applied in this research within the Bayesian framework. The imputation of missing data values in this project has incorporated majority of the attributes derived from the spatial microsimulation, which combined the LENNON tickets, Census interaction and NRTS data. Recall also the further contextual within network attributes were derived from the GIS-GTFS model.

7.6.1 General remarks

Journey factors of season tickets are missing simply because the description of the season ticket product as well as the network sensors do not pick up journey numbers for season tickets. It can reasonably be assumed that the mobility of the passengers are not constrained or affected by the presence or absence of count sensors on the network, as all passengers are assumed to be using a ticket within its validity. As such, it can be stated that the mechanism of missing values is distinct from the mobility mechanism. This is one of the pre-requisites which form the basis for object imputation of missing data values. In some transport systems, the sensors on the network record passenger journeys up until a fixed cap. A typical example is the case of TfL[89] transit (buses and trains) where fares are capped after

[89] TfL stands for Transport for London, and is a local government body responsible for the entire public transport within the 32 administrative London boroughs and City of London, an area and population of about

accumulating a number of journeys). In essence once the journey factor reaches a predefined amount, the fare charges no longer increases and stays fixed at the cap value. A missing journey factors in that instance would indicate that the passenger exceeded a minimum threshold in number of journeys. In these circumstances the mechanism for the missing data is related to the number of journeys made, in turn related to the mobility, and then the assertion is that the missing data mechanism is related to and not distinct from the mobility mechanism. Under such conditions (as with the TfL) an objective imputation cannot be developed simply from inclusion of more exogenous and endogenous attributes, as would otherwise have been required based on Rubin's postulation (Rubin, 1976). When the missing data mechanism is not distinct from the mobility mechanism, a model would have to be created for the missing data mechanism as part of any data imputation exercise. Such a model would be similar to the one developed in this research for the missing journey factor indicator (Model 3, Table 7.1.and Figure 7.3).

The imputation database consists of a total of about 160 exogenous and endogenous variables. Some of the variables are correlated, and these have been carefully selected to avoid issues reported in the literature concerning correlated covariates in MCMC algorithms (Congdon, 2014, Gelman et al., 2014a, Lunn et al., 2000, Plummer, 2003, Smith and Roberts, 1993). Lists of the variables investigated during the model development stages are included in Table 7.3. The choice of variable is sometimes driven by the efficiency of convergence of the Bayesian model, as well as the particular mobility analysis context. It was for instance observed that there were no marked advantages in using station entry/exit as opposed to residential origin/ final destination. The former however yielded better convergence in the mobility model, so was adopted unless the context of a particular case study warranted otherwise. All the variables considered during the development stages of the imputation models are listed in Table 7.3, with those variables eventually adopted highlighted by an emboldened '**Y**' in red.

The R-software created to fulfil the Bayesian imputation have been described in Appendix C, with the detailed software scripts included in the CD accompanying this thesis.

1.6Km^2 and 8.5million respectively. The TfL transport system spans the entire Greater London.

Table 7.3 List of variables considered in the models reported in the thesis.

Outcome explanatory (O) or Exposure predictor (E) variable	**Model 1: flows**	**Model 2: journey factor**	**Model 4: ticket validity**	**Model 5: group stn. ind.**
Passenger count (Freq)	**Y**			
Origin - station entry	**Y**	**Y**	**Y**	**Y**
Destination – station exit	**Y**	**Y**	**Y**	**Y**
Origin – house/ other Postcode	Y	Y	Y	Y
Destination – work/ other Postcode	Y	Y	Y	Y
Distance (journey, access, or egress)	**Y**	**Y**	**Y**	**Y**
Cost (monetary, or time)		Y	Y	
Purpose	**Y**	**Y**	**Y**	**Y**
Age/Gender	Y	Y	Y	Y
Income	**Y**	**Y**	**Y**	**Y**
Household Type/children/cars	Y	Y	Y	Y
Ethnicity	Y	Y	Y	Y
Access mode to railways		Y	Y	Y
Egress mode from railways		Y	Y	Y
Daily Rates of ticket use – journey factor	**Y**	**Y**		
Ticket validity period (days/months)	**Y**	**Y**	**Y**	
Group station flow/indicator	**Y**	**Y**	**Y**	**Y**
Number of stops	Y	Y	Y	Y
Travel time (gross/actual)	Y	Y	Y	Y
Stn. arrivela/actual departure - time	Y	Y	Y	Y
Station access/egress distance	Y	Y	Y	Y

Note: Model 3 has been excluded because it was not adopted to produce the results presented. Model 6 is excluded as it is simply the ‘dsum’ construct (included in the CD accompanying the thesis). The ‘Y’ in bold indicate the variables that were eventually used.

PART 3

This is the last part of the thesis and is made up of Chapter 8 and Chapter 9. Having harnessed the large consumer data to create an attribute rich micro-level representative population of railway passengers embedded in the wider population, two case studies using such micro-level data are presented. The case studies highlight typical use of such micro-data in the detailed spatial analysis of mobility phenomena. In the last chapter of this part, ethical issues regarding consumer big data are reviewed, discussion of the research presented, and a synopsis of the thesis is given, as well as limitations of the work, conclusions and work anticipated for the future.

Chapter 8
CASE STUDIES IN WEST YORKSHIRE

This chapter presents two case studies of urban mobility on the railways in West Yorkshire. The first case study investigates the phenomena called rail-heading, whereby passengers travel further to access the rail service when there are closer access points. The second case study investigates the effect of an intervention of a new railway station at Kirkstall Forge. The hypothesis is that the availability of representative micro-level passenger information facilitates the investigation of such mobility phenomena. Recall that the representative railways population was created by harnessing the large consumer ticketing data available from the railways using a set of complementary technologies: m-IPF spatial microsimulation, GIS-GTFS data model used for network simulation, and Bayesian imputation, as presented in the previous chapters of this thesis. The rail network in the West Yorkshire County is shown in Figure 8.1, indicating the stations, Rail Zones, and the new Kirkstall Forge railway station. The Bayesian analysis shows that rail-heading is directly effected by the geographic location of residence of the passenger, indicative that such phenomena is attributable to the service provision, and less so to the behavioural attribute of passengers. The mobility dynamics of the new station at Kirkstall Forge is adequately modelled as a geostatistical phenomenon driven by mobility activity at nearby railway stations.

The basic modelling processed adopted for the two case studies is as follows: the hypotheses are that the rail-heading phenomenon is either an individual-level, local-level, or global level phenomena. At individual-levels, rail heading would be ascribed to the particular idiosyncratic heterogeneity in behavioural attributes of each passenger. At a local-level, rail-heading could be ascribed to the socio-demographic attributes of groups of passenger, or to local geographic influences. At scales across the West Yorkshire study area, a global-level rail-heading would be enmeshed within the latent heterogeneity of societal mobility. To assess these rail-heading hypothesis the rich individual-level dataset created by harnessing the LENNON ticketing data are used[90] for mobility analysis as follows: The rail-station access

[90] The LENNON ticketing data was harnessed by combining with the Census and NRTS. The simulated population was then used as input to a GIS-GTFS data model. These yielded a requisite analysis datasets with rich sets of exogenous and endogenous attributes.

distance of each simulated passenger formed the outcome variable. The explanatory variables consisted of the entry and exit stations, the Postcode Sector of residence and destination of passengers, number of household cars, train stops along each journey, railway access mode, travel cost, journey times, and egress mode. These variables enable an investigation of the range of individual, local, and global mechanisms for rail-heading.

For the impact of a new station at Kirkstall Forge, the hypothesis is that each rail station in the study area are effectively points for sampling the mobility interaction across the rail network in the West Yorkshire study area. Each sampling produces a sampling distribution which is conditionally independent of mobility activity at other locations across the network (as such the point measurements enjoy exchangeability). The hierarchy of independence is associated with the origin and destination stations, as well as the distance between associated pairs of stations. The distance between any two stations also serves to weight the extent of relationship between stations. A hierarchical geo-statistical Kriging model is adopted to model the effect of a new rail station. The outcome variable is the count of passenger traversing between the two stations. The explanatory variables are the associated origin and destination stations, as well as the distance between the rail stations. Details of these models are presented in the methodology section of this chapter.

8.1 Introduction

Rail-heading is the name given to the phenomena in the rail sector whereby passengers are observed to regularly travel further distances to access a rail service when there are closer access points to the rail service. This propensity to rail-heading could be a result of a plethora of reasons. Passengers could travel further to cross a zonal boundary so as to benefit from lower within-zone fares. This may result in seemingly unexplained congestions on alternative modes of transport, with the attendant adverse logistics impacts (Fowkes et al., 1985, Whiteing et al., 2003), highlighting a need for changes in the rail fare structure. Rail-heading could also result from passengers wanting to optimize their utility by accessing the rail station at an entry point that provides more service frequency and fewer or more stops, for a quicker journey. Passengers could also be attracted to routes where car parking facility is available and at lower costs. Other influences could be the choice of a more scenic route, or a route that enables the fulfilment of a range of activities.

The phrase rail-heading was first used in a Fareham Borough Council report on future railway fares policy (Viccars, 2002). In a bid to make rail travel an effective alternative mode of transport, rail fares are typically structured to encourage network travel using various ticket products. Managing these ticket products often becomes complex as there are about 856 different ticket products available (as at the time of research in June 2017), with added complexity each time new products are introduced. Often, the effects of restructured fares are not always straightforward to manage by both the passengers and the rail authorities. An instance given in the Fareham report was that a restructured ticket product desirably increased rail passenger demand around Southampton and Portsmouth, but inadvertently led to a prevalence of passenger journeys by other modes to the peripheral stations of London. Rail passengers used the cheaper Southampton fares to get to within the vicinity of City of London, and to benefit from lower central London Fares. The resultant congestion and pollution at the peripheral London stations was not an anticipated desirable effect.

In the Fareham case, the response of the railway authorities was to recommend a simplification of ticket product to enable an easier study and anticipation of the rail-heading phenomena. As a general point, although the London area (Greater London) is seen as a particular case for rail (and road) where special measures might be needed, the rail-heading phenomena is not a special problem for London, but is something that is known to occur in numerous locations around the UK. As a result, in response to the Fareham case, further nationwide policies were introduced aimed at minimizing fare differentials set by different Passenger Transport Executives. Both the Fareham and nationwide policy recommendations are a simplified solution to a complex phenomenon, as rail-heading could be an individual (i.e. micro-level) phenomena. The dearth of micro-level information on passengers makes it more difficult to address individual phenomena like rail-heading. This research aims to address such dearth. A detailed spatial analysis would indicate categories of passengers that are responsible for driving rail-heading to enable tailored solutions to the problem. This section develops Bayesian spatial analysis to understand the drivers of the rail-heading phenomena in the West Yorkshire study area. The Bayesian framework developed includes a range of methodologies that have been applied to ecological, medical and epidemiological research to study the spatial variation in the risk of disease (Bivand et al., 2008, Lawson et al., 2003, Penny et al., 2011). These methods have been developed in this research to ascertain the propensity of rail-heading in spatial locations.

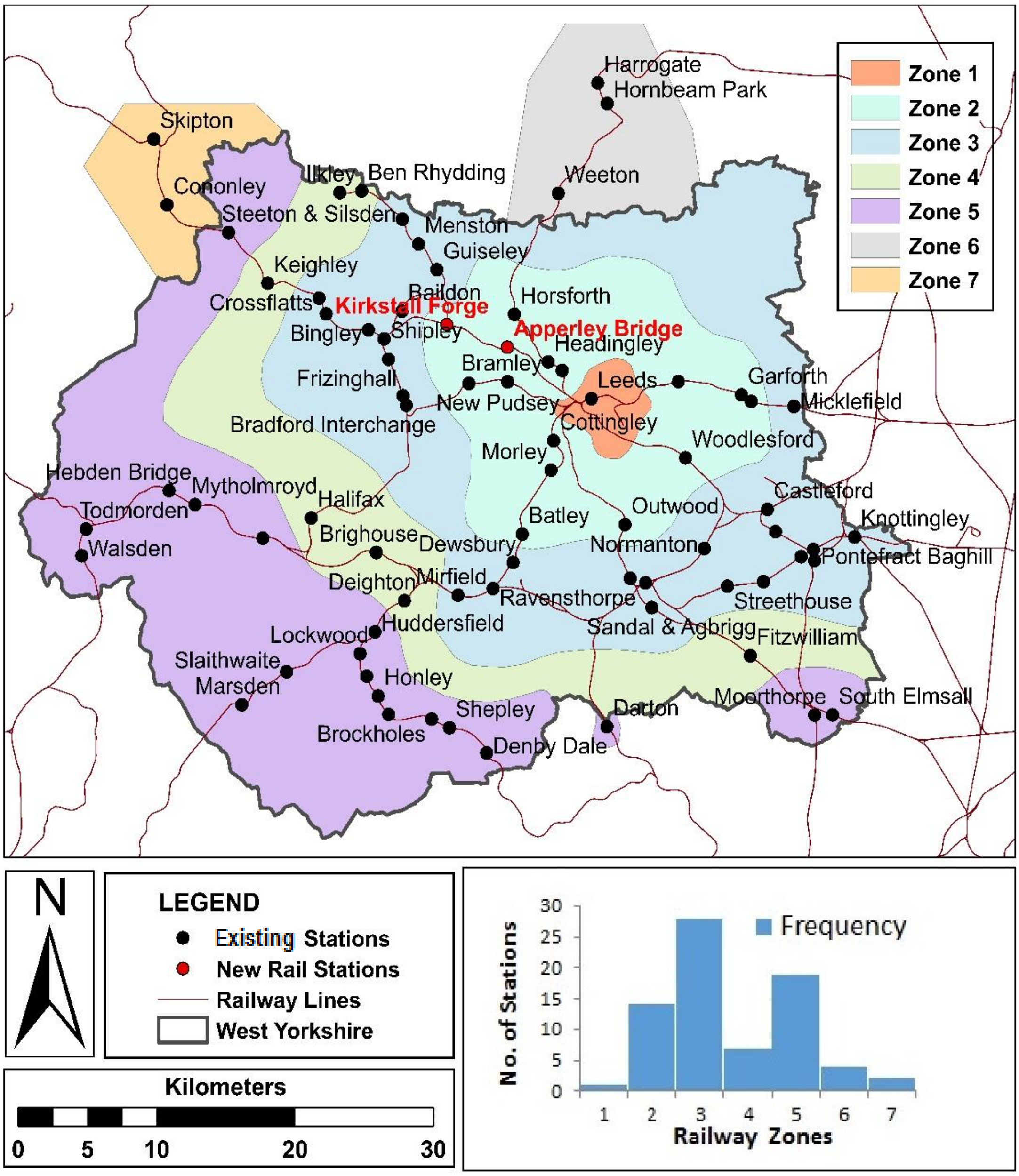


Figure 8.1 Rail network, zones, and new Kirkstall Forge station.

Another case study investigated is the micro-level demand resulting from a new railway station at Kirkstall Forge. New railway stations are typically introduced as incentives to encourage individuals to use rail as the choice mode of transport in fulfilling daily activities. Current cost-benefit analysis and impact assessments rely on gravity laws or spatial interaction models which project an average estimate of the passengers that would use the railways facility. It is typically assumed that in the UK urban setting, each railway station would have a catchment area which forms the pool of potential passengers. This is normally set at a radius of 800m (Preston, 1987, Preston, 1991), although the size of the population in the adjourning catchment neighbourhood will also influence the demand for the rail service.

This simplistic model of demand for a new railway service is useful in many settings where it is desirous to estimate aggregate demand. However, in scenarios where services are increasingly tailored to individuals, the idiosyncrasies of passenger circumstance have to be included in any demand framework. The current gravity law modelling construct is known to suffer flaws in the form of analytic inconsistencies (Simini et al., 2012), with suggestions that mobility is better modelled as a stochastic process of local mobility decisions. The availability of strategies for creating individual-level populations of railway passengers opens up opportunities for micro-level spatial analysis. The Bayesian framework inherently has the potential to construct models conceptually better at reflecting the stochasticity in individual mobility behaviour (Spiegelhalter et al., 2002). Bayesian models when applied to mobility is premised on individual movement being a chained sequence of events that are more dependent on local decisions, guided by an overarching wider activity schedule. This informs the basis for using Markov chain Monte Carlo (MCMC) models to represent a wide range of phenomena in nature (Gilks et al., 1995).

Another aspect of new train station demand not typically captured within the framework of the gravity model is that part of the demand arises from transfers from other modes of transport (Leeds, 2006), and as such any assessment of the benefits of a new rail station would require an incorporation of information from other transport modes. The strategy proposed in this thesis for investigating the impact of a new station exploits detailed endogenous and exogenous information associated with the wider population (with rail passengers embedded). A recent report commissioned by Network Rail (Gleave, 2011) indicated that a new station investment has value beyond being an access point to the railway network, but should increasingly be seen as a focal point of cities and centres of economic activity. As such, an assessment of the impact of a new station, ought to incorporate a broad range of related attributes of a simulated population, like income status of passengers, journey purpose, origin and destination locations of passengers, as well as a broad range of additional socio-economic attributes embedded in the simulated micro-level population. The Bayesian framework holds promise in this regard, providing a flexible framework for assessing a range of covariates, as well as enabling constructs to be used to better reflect the mobility scenario being modelled. This framework facilitates the incorporation of additional information garnered from experience and knowledge about the variables being modelled.

8.2 Literature review

This section reviews the concepts behind the range of methods developed to analyse spatial phenomena in this research. Two modern modelling notions are reviewed, that of exchangeability whereby the phenomena being modelled can be conceived as a hierarchy of similar events. Another notion is that of non-stationary distribution whereby the phenomena being modelled are a geographically related set of identically distributed events. These modelling concepts encompass the descriptive range of urban phenomena. These notions supersede the standard regression models which are traditionally fitted to data, but which do not aim to capture detailed spatial variation or any hierarchical aspects. These newer modelling concepts are more representative of the complexities in mobility phenomena, as mobility is a locally varying phenomena influenced by a hierarchy of events (Simini et al., 2012). The availability of novel consumer data has opened the realm for the development of more sophisticated analysis methods, buttressed by developments in computing giving the analyst the ability to easily and flexibly specify more representative models capturing relevant facets of mobility. These issues have been exasperated by the demands for more sustainable and efficient management practices, and for services specifically tailored to the demands of the individual.

8.2.1 Spatial modelling concept

Models of urban geographic phenomena are captured by representing physical objects at different scales in a geographical information system (GIS). The GIS data can be separated into two categories, vectors and raster’s (Dangermond, 1992). Raster data include imagery and pictures, but the predominant GIS data category is vectors. Typically vector modelling constructs used are points, network lines, or aerial polygons depending on the context being modelled. Points are used to represent discrete nonadjacent features where the length or area information are not relevant. For example, features like a location of interest, a school, a hospital, or a railway station can be represented by a point. Lines are used to represent linear features with only one dimension, like roads, river trails, railway lines, and streets. Aerial polygons are used to represent 2-dimensional features like the boundary of a city, a school perimeter, or a rail travel zone. The objects are typically described by sets of variables and values, and a model is constructed to capture the joint relationships between the facets of objects, variables and values.

A first basic model of mobility on the railways for example, could assume that the behaviour of passengers are mutually exclusive, and each passenger is dealt with in isolation. Such an assumption for mobility would be naïve, as such a model would be exhaustive and does not provide a summary of the mobility, and is not representative by not accounting for the interaction between passengers. A second simple but more effective model can assume that all passengers belong to the identical pool of railway travellers, and that one individual’s ticket procurement is distinct and independent of another passenger’s access to procuring a ticket. Such mobility can be captured by a global model describing the unique joint distribution of the variables describing the objects (passengers). The unique distribution would be the parametric value of the mobility model, and the observed and unobserved variables would describe the heterogeneity in passenger behaviour. Such a modelling framework has yielded the classic spatial interaction models of mobility, the gravity models, entropy maximisation models, radiation models etc. (Fotheringham, 1983, Gonzalez et al., 2008, Simini et al., 2012, Wilson, 1969). A further improvement on this second type of model could account for the non-stationarity in mobility behaviour, positing that mobility in one location could have location specific influences which die down with decrease in proximity. In such a case the dependence of the objects is due to the similarity in geographic proximity, and not due to similarity in object types.

There are fundamentally two systems for modelling such non-stationarity[91]. In the first system described above, each object is assumed to be influenced more by other objects in close geographic proximity, and less so the further away an object. In such a case, objects are related by geographic proximity. In the second system, objects are related by being similar. As such objects may not necessarily be in the same geographic proximity, but are proximal by virtue of coming from the same distribution, warranting their attributes to be similar. The first notion involving spatial proximity is adopted in developing non-stationary analysis models like the geographically weighted

[91] It is noteworthy to point out that formally, geographic non-stationarity is conceptually distinct from group non-stationarity, albeit that both are founded on the concept of objects (for example rail passengers) who are identically distributed by coming from the same wider pool of railway passengers. The distinction is that the former are founded on the concepts of object independence. The latter are premises on the concepts of groups of objects that are conditionally independent (exchangeable).

regression (GWR) typically applied to areal lattice (Brunsdon et al., 1996, Cleveland and Devlin, 1988, Fotheringham et al., 1998), and geostatistical kriging (Cressie, 1990, Stein, 2012, Lake, 2016) typically applied to point objects. The second notion of proximity of object by virtue of similarity is adopted in developing hierarchical models and mixture models (Carlin et al., 1992, Gelman and Hill, 2006, Richardson and Best, 2003) based on the notion of exchangeability (De Finetti, 1972).

In summary, three fundamental modelling systems are identified: the first system is the simple model of independent objects. The second system construes object as independent samples which are identically distributed. This second system can be further complicated by introducing geographic (or so called spatial) non-stationarity. The third system of modelling is the case where the objects are considered to belong to hierarchies of groups, whereby objects in a group are considered similar irrespective of their geography. These three scenarios are represented in Figure 8.2. The first case (on the left) represents where the objects $y_1 \dots \dots y_n$ are mutually exclusive. In the middle case, the objects are independent and identically distributed, with distribution θ. In the third case (on the right), the objects are conditionally independent, subject to the higher parameter attributes μ, σ^2. These frameworks represent cases where the objects are respectively, exclusive, identically distributed, and exchangeable as discussed briefly below.

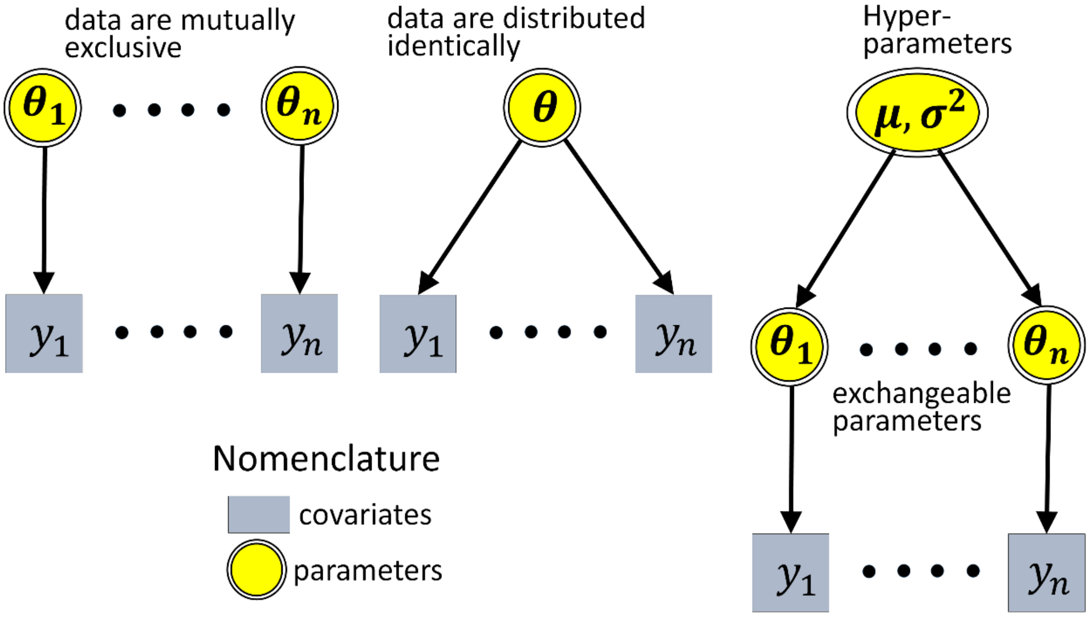


Figure 8.2 Exclusive, identically distributed, and exchangeable (hierarchical) outcome variables.

8.2.2 Exchangeability phenomena

Consider a scenario of passenger access distance to train stations. The access distance in a particular zone, say Zone 1 could simply be a result of the zone being large, and having limited transport facilities such that there is only one station in the middle of the zone. As a result, the access distance of rail passengers in Zone 1 which is on average geographically large would be relatively high compared to the zonal average access distance to a train station (i.e. considering the individual access distances to train stations of the other zones in the region). For another zone in the region, say Zone 2 of smaller geographic size, and similarly with one train station in the centre of the zone, the resulting access distances for passengers in this Zone 2 will be relatively lower than for Zone 1. Imagine there is a lattice of spherically shaped zones in the region, all with a single station in the middle of the zone. For a zone, say Zone 3 at a distance away from Zone 1 and Zone 2. The access distance to the train station in Zone 3 would not be dependent on its proximity to Zone 1 or Zone 2, instead the size of Zone 3 would decide the access distance to the train station in Zone 3. In this example, the size of geography is the unique parameter driving the railway station access distances. In this simplistic example, the access distance is assumed to be a straight line path whereby the topology of the zones is such that the road network is a non-convoluted[92] straight line. To formalize some of the concepts presented, the notion of conditional independence is introduced by way of developing the simple example of train station access distance presented so far.

Assuming that the simple example of the spherically shaped zones is complicated further, such that instead of a single train station in the middle of the zone, there are a number of train stations in each zone. Assume also that the distribution of the count of number of stations in each zone is governed by a unique parameter derived from say a Poisson likelihood function (to represent count of stations in zones). Then, the relationship between the size of geography and number of stations of the zones are independent, conditional on the parameter governing the numbers of stations in the zones. To see why this is the case, imagine that as we go farther away from the centre of the region (made up of all the zones), that the number of stations in a zone decreases. Similarly as we go farther away from the centre

[92] In the wider Yorkshire study area, the realistic topology is such that the roads are convoluted due to diversions around hills, rivers, etc.

the size of zones increases. In essence then, the parametric value is that as the size of the zones increase the number of stations decreases, the farther away from the centre. However, conditional on the parameter value (i.e. within those zones of the same size and distance from the centre), we infer that they are independent and any variation in values of these similar zones can be considered purely random. As such we say that similar zones are independent of different set of other similar zones, conditional on the parameter value.

In the case described above, the access distance to train stations across the entire region (i.e. including all the zones) is not driven by geographic proximity of the zones. Instead, it is driven by the parameter distribution across the entire region. Zones of similar parameter value have the same number of stations within. In such a case a model of zonal access distance would be better constructed by a hierarchy of aggregate sets of zones, with each aggregate set constituted by zones with statistically similar attributes. The zones within an aggregate are similar, but independent of each other. The aggregate zones in turn are conditionally independent of other aggregate zones. As such the order in which the aggregate zones are drawn is not important. The variation in the zones within an aggregate can be considered random. The order in which aggregate zones are drawn to create a parameter distribution is unimportant, and such sets of aggregate zones are described as exchangeable (De Finetti, 1972, Gelman and Hill, 2006, Lunn et al., 2012). The parameter value of the entire region is called the hyper-parameter governing the phenomena in the region. The identical distribution within each aggregate set of zones (conditional on the hyper parameter) are in turn governed by a lower level parameter distribution, which is conditionally independent of values in any other aggregate groups.

In this research, the notion of exchangeability is used to develop a hierarchical conditionally autoregressive (CAR) model of rail-heading in the West Yorkshire study area. The rail-heading model developed accounts for geographical similarity as well as similarity in clusters of phenomena. The advantage of developing the hierarchical model within the Bayesian framework is that it enable the degree of similarity across the various characteristics of mobility to be quantified (for example passengers travelling from the same location or for the same purpose can be classed as being within the same hierarchy, or the hierarchy may be constituted by income bracket, or even age group), and this is useful in conducting studies on drivers of mobility phenomena.

8.2.3 Identically distributed and non-stationarity

Whilst the notions of conditional dependence and exchangeability presented in the previous section address hierarchical mobility models that represent clusters of phenomena that each are constituted by similar events, the notions of identical distribution and non-stationarity, are important in conceptualising mobility whereby the proximity of locations imply proximal phenomena and dependency. A major rail station (MS) of high mobility activity for instance could be located between two smaller stations (called SS1 and SS2). The distance from the MS to SS1 could be thrice the distance from MS to SS2. However, if for instance mobility is driven by a station's catchment area socio-demographic profile, then it is plausible to think that SS1 could have higher mobility activity than SS2 despite SS1 being farther away from MS. The conventional notion however, has been to assume that locations in the same vicinity behave similarly, and as such a kernel can be defined to express the relatedness (and non-stationarity) of geographically proximal locations. Such a framework is a development of the concept of independence and identical distribution as illustrated by the sketch in the middle of Figure 8.4. In such a framework the constituent parts (which could be points, lines, or areas) are assumed to be identically distributed thereby enabling a global average to represent the heterogeneity in such phenomena. In the case of mobility on the railways, these units are passengers in mobility interaction on the network. In modelling the mobility on the railways in this instance, the conventional assumption would be that the data are independent and identically distributed (iid) in the sense that the passengers arise from the same (identical) population of railway travellers, and also that one individuals ticket procurement is distinct (independent) of another passengers access to procuring a ticket.

To extend this iid concept to account for geographic non-stationarity, kernels are defined postulating that points in close geographic proximity are similarly behaved. The kernel density function effectively weights the influence on phenomena more towards geographically proximal aspects. This concept has been used in the literature to create the relatively established methods like the geographically weighted regression (GWR) (Brunsdon et al., 1996, Fotheringham et al., 1998, McMillen, 2004). In the GWR, phenomena in a geographic vicinity are assumed to be driven more by aspects in the vicinity, and less so by more distant aspects. The density of the kernel decides the nature of the influence. The concept of GWR has had application in transport to explore spatial variations in access to different modes of transport

(Andersson, 2017), and to develop planning tools for estimating predictors of collisions on transport networks (Hadayeghi et al., 2010). The notion of GWR is conceptually similar to Poisson-gamma moving average models (Ickstadt and Wolpert, 1997) which express the relationship between areal Poisson counts as a convolution or moving average. The similarity in GWR and Poisson-gamma models is the notion that phenomena in an area location are related to that at other locations in the same geographic vicinity.

The non-stationarity in areal phenomena can be extended to point phenomena, such that geographically proximal points are more similar and related than points at a distant geography. Such concepts have been developed in the field of geo-statistics (Cressie, 1990, Stein, 2012), dealing with spatial-temporal non-stationarity. Whilst GWR's are applicable to areas, point phenomena are similarly modelled using geostatistical kriging methods (Cressie, 1990, Lake, 2016, Stein, 2012), which assume that phenomena in the vicinity of points are a weighted function of the phenomena at the individual points. In kriging, instead of defining a kernel to weight the influence of points on the mechanism being modelled, an interpolation enables the deduction and imputation of unobserved point values. Whilst GWR is aimed at area activity, kriging is aimed at point phenomena.

Kriging has been used in the literature to estimate radioactive contamination at different point locations on the Island of Rongelap (Diggle et al., 1998). A similar concept is developed in this research to assess the influence of a new railway station on the mobility interaction on the railway network. The railway network is seen as a hive of mobility interaction, and each railway station is seen as a sampling point on a network of continuous mobility interaction. The activity measured at each station represents the sampling distribution of mobility taken from a point on a continuous network. On this basis, a geostatistical kriging model is developed to impute phenomena at the location of a new station. The Bayesian framework is chosen for this analysis, as the kriging model developed is hierarchical, and such hierarchical models can be implemented only in the Bayesian framework. Double constraints that ensure balancing of the origin and destination demand populations on the rail network can be embedded within a larger graphical kriging model in the Bayesian framework (Escobar and West, 1995, Roeder, 1990).

8.3 Methodology

A conditionally autoregressive (CAR) model is developed for use in investigating the rail-heading phenomena within the West Yorkshire study area. In principle, the CAR model specifies how phenomena in one geographic area L_i relates to phenomena at another area located within area L_j, where $i,j \leq N$, the number of areas in the geography being analysed. Weights are specified to express the spatial dependence between locations within areal or lattice data. A range of CAR models are implemented in the literature (Lee, 2013, Lunn et al., 2012). Information about the geographic boundaries and spatial adjacency of the areal data are imported from a GIS into a Bayesian model (Lunn et al., 2000, Plummer, 2003, Spiegelhalter et al., 2007, Sturtz et al., 2010). A univariate conditional distribution is then created to relate covariates in one geographic location to those at other locations. The Bayesian CAR model is enhanced to account for mobility associated with not only geographical non-stationarity, but also non-geographic (so called unstructured) hierarchical phenomena.

Secondly, a geospatial kriging model is developed to ascertain the influence of a new railway station at Kirkstall Forge West Yorkshire. Vectors of random variables P_i and P_j associated respectively to point locations i and j, (where $i,j \leq N$, the number of points in the geography) are related via a specified multivariate distribution. The multivariate covariance is then specified as a function of the distance D_{ij} between the points. The parameters of this function are deduced to enable an estimate at a new unobserved location representing a new rail station. The kriging model developed in this research is hierarchical to better reflect the range of clustered phenomena associated with a new railway station. This model is the first application in the literature of such a hierarchical kriging model. The analysis has been illustrative, highlighting the Bayesian development of these methodologies.

It is noteworthy to point out that the ability to perform Bayesian modelling to understand drivers of mobility in these case studies, is a result of the availability of synthetic, attribute-rich, and representative micro-level data created from a set of complementary methodologies developed in the previous chapters. The methodologies are: spatial microsimulation of skewed seed data, GIS-GTFS network simulation, and Bayesian imputation. The attribute-rich dataset thus facilitate the CAR model which is developed to undertake the rail-heading case study, whilst a geostatistical hierarchical kriging model is developed for the second case study which investigates the effect of the new station at Kirkstall Forge.

8.3.1 CAR-BYM model of rail-heading

This section develops the rail-heading analysis methodology within West Yorkshire County. The analysis excludes flows that emanate from and terminate outside the County, as these flows are not available to the project[93]. A conditionally autoregressive (CAR) spatial model (Fotheringham et al., 2003, Lee, 2013, Lunn et al., 2012) is used to study the within and across areal-level variability in individual-level access distance to the rail service. This rail-heading methodology developed in this research is similar to that adopted in the literature for areal-level disease mapping (Besag et al., 1991, Kelsall and Wakefield, 1999, Richardson and Best, 2003). The CAR method uses an area-level hierarchical random effect (Gelman et al., 2014b) to smooth the individual-level station access distance. An additional unstructured random effect $R_u[i]$ is added to the CAR model to capture any effects of unobserved unstructured hidden (latent) covariates (Lunn et al., 2012). Within the Bayesian model, $R_u[i]$ takes on a prior distribution typically a normal distribution, and Bayesian learning enables the exact form of such a distribution to be deduced by identifying its parameter values. The formulation that includes hierarchical random effects is called the BYM model (Besag et al., 1991), which is named after the authors who made the original model proposals. As such, the rail-heading model developed in this research is called the CAR-BYM model, and this model is used to deduce the area-level and non-area-level drivers of the rail-heading phenomena.

In essence, the passenger access distance within the database is regressed against the socio-demographic, and other exogenous and endogenous covariates. If for instance a particular age or income bracket results in higher access distance, then all other variables being constant, the inference is that passengers in the particular age or income category have a higher propensity for travelling further to access a rail service. Similarly, if passengers with lower travel times for distance travelled regress better with access distance, then the inference is that rail-heading is driven by a desire of that group of passengers for time utility optimization. If particular origin-destination (OD) flows are associated with an increased access distance,

[93] The impact is that the rail-heading phenomena reported would not account for those passengers crossing the West Yorkshire zone 5, 6 and 7 boundaries (see Figure 8.1) to benefit from cheaper rail fares. Further, as discussed in Chapter 5 and Chapter 6 the comprehensiveness of passenger volumes may be compromise, however, this effect would be nominal due to the representativeness of the simulated population.

then perhaps that flow is poorly served and passengers have to travel further to access that entry station. A summary of the nascent aspects of the R-software code for implementing the CAR-BYM multivariate model for rail-heading phenomena is included in Figure 8.3, and reproduced in Appendix E.1 with model annotations. Details of the CAR-BYM model implemented in open BUGS (Spiegelhalter et al., 2007, Sturtz et al., 2010) software are included in Appendix E.2. and Appendix E.2.

```
model {
for (i in 1 : N.zones) {
        for (k in 1 : Z) {
        Y[i, k] ~ dpois(mu[i, k])
        log(mu[i, k]) <- log(E[i, k]) + alpha[k] + S[k, i] +
                                              U[i, k] + beta[i]*(X[i, k])
        X.pred[i,k] <- beta[i]*(X[i, k])                         }
        beta[i] ~ dnorm(mu.beta, inv. Ω.beta.sq)                 }
# priors for structural
        S[1: Z, 1 : N.zones] ~ mv.car(adj[], w[], num[], Ω[ , ])
        for (i in 1:sumNumNeigh) { w[i] <- 1 }
        for (i in 1 : N.zones) {U[i, 1 Z] ~ dmnorm(zero[], tau[ , ])}
        for (k in 1 : Z) { alpha[k] ~ dflat() }
# precision  and covariance matrices of MVCAR
        Ω[1 : Z, 1 : Z] ~ dwish(R[ , ], Z)
        sigma2[1 : Z, 1 : Z] <- inverse(Ω[ , ])
# precision and covariance matrices of MV Normal
        tau[1:Z, 1:Z] ~ dwish(Q[ , ], Z)
        sigma2.U[1:2, 1:2] <- inverse(tau[ , ]);
# priors for non-structural
        mu.beta ~ dnorm(0, 0.0001)
        inv. Ω.beta.sq <- 1/ Ω.beta.squared
        Ω.beta.squared <- pow(Ω.beta, 2)
        Ω.beta ~ dunif(0, 100)
        sumvar  <- sd(S[1, ])*sd(S[1, ]) + sd(U[ ,1])*sd(U[ ,1]) +
                                              sd(log(E[ , 1]))*sd(log(E[ , 1]))
# results for investigation
        pS          <- sd(S[1, ])*sd(S[1, ])/sumvar
        pU         <- sd(U[ ,1])*sd(U[ ,1])/sumvar
         pX         <- sd(X.pred[ ,1])*sd(X.pred[ ,1]) /sumvar
        pE          <- sd(log(E[ , 1]))*sd(log(E[ , 1])) /sumvar
} end of script

# typical initial conditions
        List (
        alpha          = c( , ),
        beta           = c(rep(0, (N.zones))),
        Ω              = structure(.Data = c( ,  ,  ,  ),.Dim = c(2,2)),
        Y              = YY, S = SS, U=UU )
# end of initial values
```

Figure 8.3 CAR-BYM multi-variate rail-heading R-script.

8.3.2 Kriging model for new rail station

The conditionally autoregressive models developed for investigating rail-heading would not be appropriate for investigating the effect of a new station. This is because whilst the incidence of rail-heading can typically be associated with a zone or a particular type or group of passengers with specific attributes, a new train station is appropriately associated with the train station as a point feature with many inherent attributes and characteristics. Mobility activity observed at a railway station is conceived as a sampling distribution at a specific point on a continuous network of passengers in mobility. As a result the rail passenger mobility related to a new station is modelled using geostatistical models which have origins in kriging methodologies (Cressie, 1990, Lake, 2016, Stein, 2012).

The development of such a geostatistical model is achieved by applying methods of Bayesian inference using the Gibbs sampling methodology (Kruschke, 2014, Lawson et al., 2003, Lunn et al., 2012, Plummer, 2003, Spiegelhalter et al., 2007, Sturtz et al., 2010). In the Bayesian framework, the vector of passenger attributes associated with each point location train station is modelled as a multivariate normal distribution. The covariance matrix of this distribution is a function of the distance between the points to reflect that correlation between geographically proximal points. The BUGS implementation ensures that the covariance matrix is symmetric and positive definite to be conceptually consistent with the Cartesian spread and correlation coefficients of the station points. The correlation function is defined to linearly decrease with distance, to a pre-defined distance of no correlation.

In the Bayesian kriging model, a new station is implemented by including an unknown point at a defined geographic coordinate. The rail mobility associated with this point is then deduced by interpolation and derived from the posterior predictive distribution (Albert, 2009, Box and Tiao, 2011, Gelfand et al., 1990, Gelman et al., 2014a, Kruschke, 2014, Lunn et al., 2012, Plummer, 2003). Two kriging models have been developed. The first model shown in Figure 8.4 aggregates the flows associated with each station. In the second model shown in Figure 8.5, the station inward and outward flows are disaggregated into components to established the contribution of each of the other stations to the aggregate inflow and outflow volumes recorded at the new station and vice versa. The second model is a hierarchical model on the origin and destination stations. Detailed R-script specifications and annotation of these models are included in Appendix E.2.

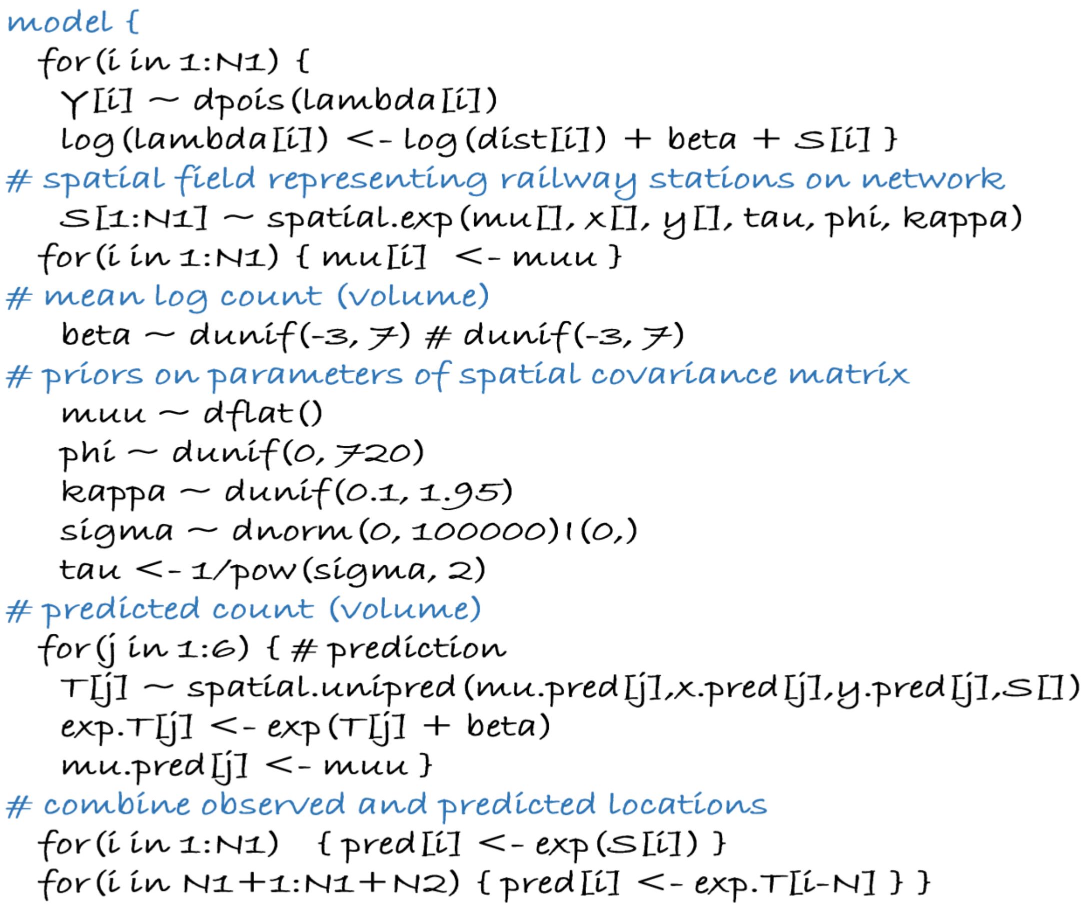
This geostatistical model developed to infer the impact of a new station is similar to models implemented in the literature (Diggle et al., 1998), and tend to produce results that are sensitive to the choice of priors. As such, guided by an example of estimating radioactivity at point locations on an island (Diggle et al., 1998), a sensitivity analysis to choice of priors was included as part of the model development process, as discussed next.

As discussed in Section 7.4.1, during Bayesian modelling, a likelihood function is created to model the outcome. Similarly a prior is specified to model the parameters. Priors can broadly be distinguished into two categories: informative or vague. In practice, vague priors tend to be used in simple situations to show the objectivity of the model, in the sense that the prior had a minimal influence on the model results. In these simple situations, the analysts has an intuition about the solution, and a uniform distribution in the neighborhood of the solution would present as a vague (also called uninformative) prior. In most practical situations, informative priors tend to be used as there is limited knowledge about the parameter results, and a tendency for sensitivities to choice of prior. In such situations the analyst exploits any knowledge of the data generation mechanism in specifying an appropriate prior to garner a solution to the model.

The sensitivity analysis chooses a distribution for the prior, as well as the parameters for this distribution. If for instance, a mechanism being modelled is conceived to consist of a phenomena represented by a normally distribution. Then if the phenomena further contains a lot of legitimate outliers, it would make sense to use a t-distribution as the prior, simply because the t-distribution has thicker tails than the normal distribution, and this tail would better contain and model those outliers (Lange et al., 1989, Peel and McLachlan, 2000). Further, having chosen the particular distribution to model the prior (i.e. normal, t-distribution, gamma, beta, etc.), there is another challenge of specifying values[94] for the parameter of this distribution. If a normal distribution prior was chosen for instance, then the mean and variance (μ, σ) of the distribution would need specifying. The sensitivity analysis as such involves iterating through the range of distributions potentially suitable for the prior, as well as iterating through the range of parameters of this distribution.

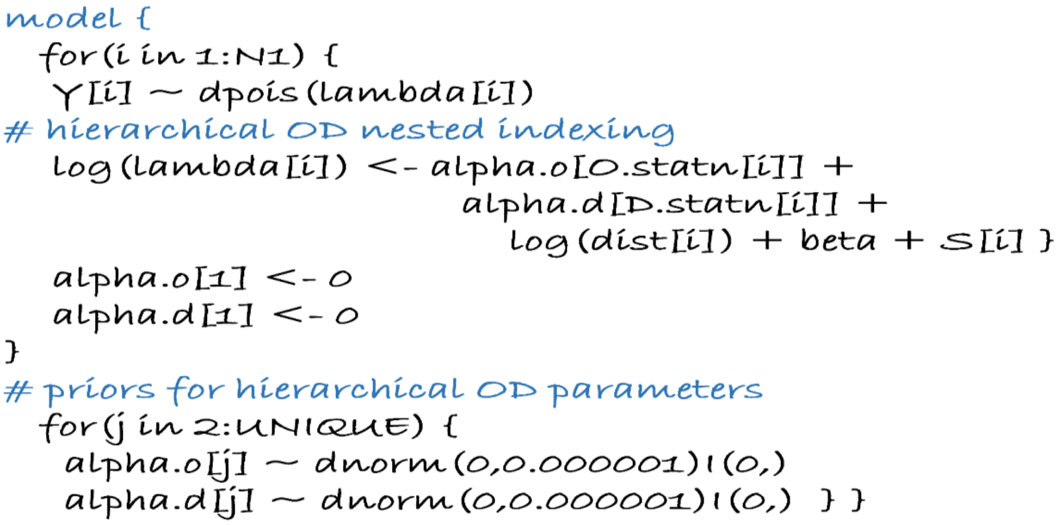
[94] Note that in Bayesian modelling, the values of the prior distribution could in turn be a distribution. This would depend on the complexity of the model being constructed by the analyst.

In the case of the geostatistical model for investigating the effect of a new station, the priors adopted were uniform distributions, as other distributions failed for the wide range of initial values tested. The sensitivity then involved iterating through a range of values of parameters (minimum, maximum) of the uniform distribution, and a range of initial conditions until convergence was achieved for the Bayesian model.



```
model {
  for(i in 1:N1) {
   Y[i] ~ dpois(lambda[i])
   log(lambda[i]) <- log(dist[i]) + beta + S[i] }
# spatial field representing railway stations on network
   S[1:N1] ~ spatial.exp(mu[], x[], y[], tau, phi, kappa)
  for(i in 1:N1) { mu[i]  <- muu }
# mean log count (volume)
   beta ~ dunif(-3, 7) # dunif(-3, 7)
# priors on parameters of spatial covariance matrix
   muu ~ dflat()
   phi ~ dunif(0, 720)
   kappa ~ dunif(0.1, 1.95)
   sigma ~ dnorm(0, 100000)I(0,)
   tau <- 1/pow(sigma, 2)
# predicted count (volume)
  for(j in 1:6) { # prediction
   T[j] ~ spatial.unipred(mu.pred[j],x.pred[j],y.pred[j],S[])
   exp.T[j] <- exp(T[j] + beta)
   mu.pred[j] <- muu }
# combine observed and predicted locations
  for(i in 1:N1)   { pred[i] <- exp(S[i]) }
  for(i in N1+1:N1+N2) { pred[i] <- exp.T[i-N] } }
```

Figure 8.4 Aggregate kriging script for new Kirkstall Forge.



```
model {
  for(i in 1:N1) {
   Y[i] ~ dpois(lambda[i])
# hierarchical OD nested indexing
   log(lambda[i]) <- alpha.o[O.statn[i]] +
                             alpha.d[D.statn[i]] +
                               log(dist[i]) + beta + S[i] }
   alpha.o[1] <- 0
   alpha.d[1] <- 0
}
# priors for hierarchical OD parameters
  for(j in 2:UNIQUE) {
   alpha.o[j] ~ dnorm(0,0.000001)I(0,)
   alpha.d[j] ~ dnorm(0,0.000001)I(0,)  } }
```

Figure 8.5 Disaggregate kriging script for new Kirkstall Forge.

8.4 Results

The result of investigating the phenomena of rail-heading and the impact of a new rail station at Kirkstall Forge are presented below. The mechanism of the rail-heading is presented. The rail-heading reported is restricted to phenomena arising from transit travel limited to within the West Yorkshire study area, as such rail-heading that might be associated with rail zone 5, 6, and 7 passenger mobility would not have been captured. The phenomena of rail heading is most pronounced in Postcode Area BD1 and in particular for passengers who started their journeys within Postcode Sectors BD12, BD11, BD15,and BD71. These Postcode Sectors are in the vicinity of the Bradford BR group of stations (i.e. Bradford Forster Square and Bradford Interchange). The rail-heading is associated with passengers who make lengthy trips of about 29km from the neighbourhood of Bradford BR to the Wakefield group of stations (i.e. Wakefield Kirkgate and Wakefield Westgate) to access the train station. These make the 29km trip by alternative modes of transport in order to fulfil a relatively shorter train trip of about 17km from Wakefield to Leeds.

In the case of the new station at Kirkstall Forge, the aggregate colume of passengers associated with the new station, and the mobility interaction between the new station and all the other existing stations on the rail network are estimated. These results are presented in the form of an origin-destination (OD) map. The advantage of estimating the new station flows using the hierarchical kriging model is the ability to deduce the disaggregate composition of these passengers, enabling the rail industry to use such information to estimate the impact of the new station on various socio-demographic groups, or which residential and work related areas would be most impacted by such a new station. The flows estimated from the kriging model were validated by comparison with MOIRA flows, and recent flows published by the ORR now that the Kirkstall Forge station has gone on stream since June 2016. The flows estimated from this research are of an order of magnitude (~45%) less than those projected by ORR. This was anticipated, as the flows emanating and terminating outside West Yorkshire (between 25%-45%) were not included in the analysis. In addition, the PTE tickets were not included in the analysis and these accounts for between 25%-35% of county rail flows. Further, it has not been possible to validate the disaggregate passenger volumes deduced from the Kriging, as there are no available datasets that disaggregate passenger flows between railway station by socio-demographic or other attributes of passengers.

8.4.1 Rail-heading phenomena

Results for the propensity for rail-heading in the West Yorkshire study area are presented. The coupled CAR with BYM geospatial model captures the within and across-area difference in passenger access distance to train stations, and this in turn is a measure of the propensity for rail-heading.

8.4.1.1 Variation due to latent random effects

The ratio of the variation due to structural-areal versus non-structural latent random effect reveals whether rail-heading is a geographic zonal phenomena, or embedded in the heterogeneity of passenger socio-demographic attributes or urban morphology. The covariates of mobility: travel distance, cost, household car ownership, train station access and egress mode, number of stops, etc., as well as a number of socio-demographic attributes are individually tested to ascertain any direct effect or relationship with access distance to train stations.

The results shown in Table 8.1 give a better insight into the mechanism of rail heading. In summary the results show that the total travel distances that passengers within a particular Postcode Sector boundary are willing to endure, relates directly to how further afield they would typically be willing to travel to access the rail service. The more household cars passengers own, the more willing they are to involve in rail heading. Rail-heading passengers are driven by a requirement to take fewer stops to get to a commensurate journey, perhaps accessing the train service at stations where they can benefit from direct trains (or trains with fewer stops) to get to their final destinations quicker. Passengers that partake in rail-heading are more likely to access the station by mode of car and less likely by mode of walking.

The R-software model in Figure 8.3 was developed to generate the results presented in Tables 8.1. The CAR component of the CAR-BYM model specifies the relationship between variables at one area (say a Postcode Sector) and those at all other areas, using conditional regressions (Besag et al., 1991). The BYM component of a CAR-BYM model imposes Bayesian learning about the strength of areal-specific spatial dependencies, by including a non-areal random distribution to capture any unobserved unstructured (i.e. non-areal) random variations. The simulated population includes information on the distance each passenger travelled to access a train station, the passenger’s origin Postcode (typically their residence) and final destination Postcode, entry and exit train station points, as well as a plethora of endogenous and exogenous passenger attributes.

Rail-heading mapping involves producing Postcode-level summaries of distance travelled to access train services. The model developed is shown in the pseudo-code in Figure 8.3. Typically, spatial dependencies are modelled using hierarchical spatially structured random effects, with the variables assumed conditionally independent given the random effects (Clayton and Kaldor, 1987, Kelsall and Wakefield, 1999). In the case of rail-heading the outcome variable $Y[i,k]$ are passenger counts in Postcode Sectors, with a Poisson likelihood function (seen in Figure 8.3). The covariate $E[i,k]$ is a measure of the expected count of passengers deduced by including the perimeter (or area) of the Postcode Sector. The covariate $X[i,k]$ represents each variable that was separately adjusted for (see list of covariates in Table 8.1). The travel distance adopted is a ratio: the numerator is the passenger distance travelled to access the train station. The denominator is the expected passenger travel distance, worked out from the distance between usual residence and nearest train station, calculated along the road network. The term $alpha[k]$ is an intercept term which represents the baseline distance travelled by passengers.

In the spatial mobility data, the sampling variability would often obscure any systematic between-area differences in access distances to train stations. A remedy would be to use a spatial hierarchical model by way of an area-level random effect $S[i,k]$ with a normal distribution CAR prior. As such, $S[i,k]$ captures the effect of unobserved spatially structured covariates (i.e. those related to areal phenomena). To capture the effect of unstructured unobserved covariates, an additional random effect $U[i,k]$ is attached to the CAR model. The $U[i,k]$ is defined as a normal distribution random effect with posterior parameters revealing any true unobserved non-structural variation. A Bayesian model then identifies the extent of the structured (i.e. spatial) and non-structural data variation by computing measures (i.e. pS, pU, pX and pE see Figure 8.3 and Table 8.1) of the posterior proportion of total within and between-area variation explained by each random effect.

The relative magnitudes of the measures pS, pU, pX and pE enable an inference of how much the propensity for rail heading is ssociated with the fact that passenger journeys originate from different Postcode sectors, or whether it is just a function of the size of the different Postcode Sector geographies. These measures also infer whether rail heading is mainly associated with other unobserved heterogeneity between passengers across the West Yorkshire study area.

As observed from the pseudo-code in Figure 8.3, pS is a measure of the influence on rail-heading of random variations in area-level access distances to the rail network. In essence, it answers the question: how does the each Postcode Sector of origin of passenger journeys affect their propensity to rail-heading? Then, pU is a measure of the influence on rail-heading of the heterogeneity in passenger mobility. Note that pU is not associated with any Postcode Sector influence on rail-heading, and in essence pU andswers the question: how do the random differences not associated with the Postcode sector where a passenger starts their journey affect their propensity to rail-heading? Note that pS and pU are divided by '$\boldsymbol{sumvar}$' (see the pseud-code in Figure 8.3). The $\boldsymbol{sumvar}$[95] is an aggregate measure of the standard deviation associated with the covariates $S[i,k]$, $U[i,k]$ and $E[i,k]$. Within the pseudo-code of Figure 8.3, the pX is the measure for the particular variable that was adjusted for (i.e. travel distance, number of cars, journey time, etc. as listed in Table 8.1). Finally, pE is (as seen in the pseudo-code of Figure 8.3) is a measure of the standard deviation (variation) in Postcode Sector sizes in the West Yorkshire study area.

The standardization is uniformly applied in computing the measures pS, pU, pX, and pE. In essence all the measures are divided by '$\boldsymbol{sumvar}$'. The division has no influence on the relative values reported for pS, pU, pX, and pE. The standardization is however included here to enable the comparison of this model with any similar developments in the literature (Besag et al., 1991, Lunn et al., 2012). The results reported in Table 8.1 as such enable the relative comparison of the values pS, pU, pX, and pE, and their absolute magnitudes are less significant. QR80 is the 80% quantile ratio used to quantify the 10% and 90% spread in specific random effects. QR80 is related to the average spread in the passenger count variable after considering the difference in size of Postcode Sector associated with each count.

Further details of similar models to the CAR-BYM are in the literature (Lunn et al., 2000, Plummer, 2003, Sturtz et al., 2010). Spatial models are usually designed to quantify variation over a 2-dimensional space, and are widely used in environmental sciences, image processing, and medicine (Lawson et al., 2003, Penny et al., 2011, Ripley, 2005). The development presented in this thesis is the first application to investigate rail-heading phenomena and mobility interaction on the railways.

[95] This value is used for standardization to enable comparison of the results with similar endeavours in other diverse fields of research.

Table 8.1 Structural and non-structural rail-heading.

Covariates	*pS*	*pU*	*pX*	*pE*	*QR80*
Travel distance	0.65	0.35	0.97	0.04	10.44b
No. of cars	0.63	0.37	0.87	0.02	1.34b
No. of stops	0.58	0.42	0.86	0.02	98.7m
Access mode	0.55	0.45	0.93	0.04	12.3m
Travel cost	0.53	0.46	0.93	0.04	1.49m
Journey time	0.50	0.50	1.04	0.04	5.82m
Income	0.47	0.53	1.03	0.06	226.8k
Egress mode	0.45	0.55	0.88	0.04	576.8k
Age	0.37	0.63	1.04	0.06	167.9k
Residual	0.29	0.71	0.00	0.06	150.46

pS **is the measure of variability due to structural random effect.** ***pU*** **is due to non-structural latent random effect.** ***pX*** **is the proportion of the variance explained by the individual covariates, and** ***pE*** **is the proportion explained by the difference in perimeter/area of the zones. ‘k’, ‘m’, and ‘b’ represent thousand, million and billion respectively.**

8.4.1.2 Propensity to rail-heading

The local propensity to rail-heading at Postcode Sector level is shown in Figure 8.6. The simulated population shows large numbers of passengers (the darker shaded areas in the proximity of Bradford stations shown in Figure 8.6) travelling as far as Wakefield stations to access the network en-route to Leeds station. The daily timetables reveal that on average a journey from Bradford stations to Leeds takes longer by rail, than a journey from Wakefield stations to Leeds (This is probably because the Leeds to London high speed trains[96] all stop at Wakefield, so the journey is around 15 mins). These intercity services are also more reliable than the local network trains and less subject to high overcrowding). Looking at the map in Figure 8.6, the darker areas (indicative of rail-heading) around Brighouse would be individuals who would actually use the Huddersfield to Leeds rail line, but

[96] The current high speed train service between Leeds and London Kings Cross has an average journey time of 2.28 hrs, and the quickest journey takes 1.98 hrs. The weekday service is provided 32 times daily. Journey times tend to be longer on weekends and holidays typically due to heightened logistical issues and passenger congestion.

that is part of the Transpennine[97] Express service and extremely unreliable, particularly in winter. The provision of access roads from Bradford to Wakefield perhaps also serves as a further incentive to rail-heading. This is perhaps further exaggerated by passengers travelling from railway zone 4 to zone 3 during their journey, thereby also benefitting from cheaper fares.

Some peripheral Postcode Sectors also reveal higher propensity to rail-head, like the Oldham Postcodes towards the west extreme of the County (with the Todmorden and Walsden stations within). However, any pronounced boundary rail-heading would be attenuated because the analysed flows have been restricted to those that emanate and terminate in West Yorkshire. To better reflect the rail-heading phenomena from access distance, the zonal areas have been included in the model as an offset to normalize for typically larger numbers of passengers associated with large Postcode Sectors.

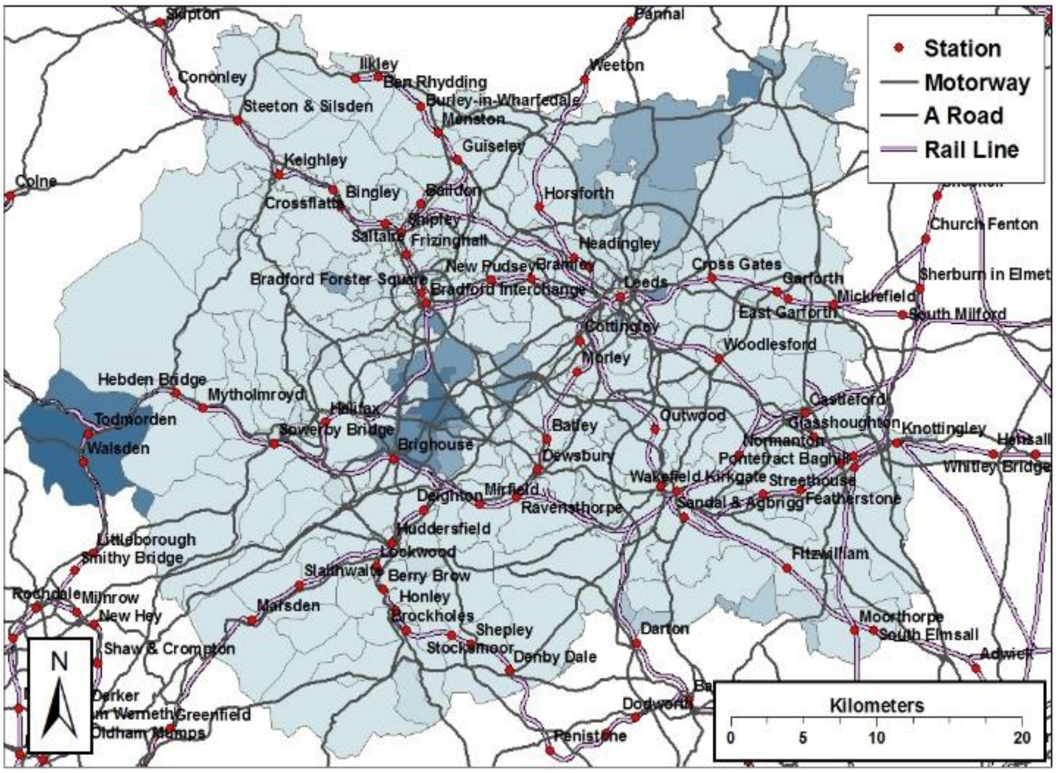


Figure 8.6 Rail-heading propensity in West Yorkshire.

[97] The TransPennine Express is a UK TOC, which runs regular express regional railway and intercity services, between the major cities of Northern England and Scotland, 24 hours daily, including through New Year's Eve night. TransPennine Express also runs a popular daily service between York, Leeds and Manchester Airport (also running at least every 3 hours at night).

8.4.2 New Kirkstall Forge station

Each rail station is conceptually assumed to have the propensity to attract and to produce passenger volumes in the spatial interaction paradigm. Kriging has been used to estimate the aggregate volume of outflows from the new Kirkstall Forge station. Similarly, the volumes of outflows from Kirkstall Forge to all other stations in the West Yorkshire study area have been estimated. These results are presented below in the form of a 3D origin to destination (OD) matrix. Apart from Kirkstall Forge, further comparable results were derived for the relatively new Apperley Bridge[98] station. The results presented in this section have not been intended to be exhaustive, but primarily serve to illustrate in a practice oriented way the Bayesian framework for estimating the impact of a new train station. The particular advantage of the Bayesian method is that the estimates can be disaggregated to estimate the impact of the new station on other existing stations, and further to deduce the category of passengers most impacted by the new station, or the type of customers an anchor tenant at the train station might be expected to encounter.

Due to the limitations in data available to the project the flows analysed in this case study have been restricted to those that emanate from and terminate in the West Yorkshire study area. About 25-35% of flows in the West Yorkshire study area have been reported in the literature to be associated with flows emanating and terminating outside the county. This is anticipated to affect the volume of passengers deduced for the new Kirkstall Forge station. It is anticipated that the constitution of the passengers would not be affected adversely as the distribution of the simulated population used for analysis was representative.

Further, tickets sold by regional integrated transport operators (the so called PTE tickets) have been excluded from the analysis as these PTE tickets have not been available to the project. As reported earlier in Chapter 2 and Chapter 3, the PTE tickets could account for about 25-35% of flows within the West Yorkshire study area, and this will affect the passenger volumes deduced for the new Kirkstall Forge station.

[98] Apart from Kirkstall Forge station which opened in June 2016, the Apperley Bridge railway station is also a relatively new station which opened in December 2015, and is along the same stretch of rail-line as the Kirkstall Forge station.

8.4.2.1 Aggregate Kirkstall Forge flows

The results shown in Figure 8.7 and Figure 8.8 are typical results from the Bayesian modelling. The results in Figure 8.7 are results for Kirkstall Forge station, while the results in Figure 8.8 are for Apperley Bridge station. These results have been generated using the model scripted in Figure 8.4, and are posterior estimates of the aggregate volumes of passengers associated with the new railway stations (i.e. Kirkstall Forge and Apperley Bridge). Bayesian results typical of those in Figure 8.7 and Figure 8.8 have been discussed in detail in Chapter 7, and reference is hereby made to that discussion.

Recall that the trace plot which is usually made up of a number of chains (in this case 3 chains) indicates whether the MCMC has reached a stationary distribution and as such converged. As seen in Figure 8.7 and Figure 8.8, the upper left panels show the trace plot which exhibits adequate mixing of the chains (after a burn-in of 10,000). As seen, there are no systematic trends as the traces progress along the horizontal axis, and this is indicative that the influence of the initial trace position has been overcome. Visually also, the three chains are evenly distributed about the mean parameter value (~8.6 in Figure 8.7 and ~8.1 in Figure 8.8). The even scatter in the plots, the overlaid nature of the traces, coupled with no systematic trends or influence of initial conditions, are indicative of MCMC chains that are representative of the posterior target distribution and that they have converged.

The next set of plots in the top right of Figure 8.7 and Figure 8.8 are the autocorrelation plots. A unique characteristic of MCMC sampling is that, if suitably designed, a sample at a point on the chain is only dependent on the immediate previous sample, and on no other prior samples. As such a test for the adequacy of the MCMC is to find out the autocorrelation between the samples with lag. As seen in the top left plots in Figure 8.7 and Figure 8.8, the autocorrelation plots indicate that the MCMC chain is highly uncorrelated indicative that subsequent samples from the chain would be independent, and effective (and efficient) for use in posterior estimates. Recall the variability and independence in values deduced from subsequent steps of the MCMC chain are indicative of the volume of parameter space explored. The more space explored, the more accurate the results, and this is related to the volume of effective samples. The results in the upper-right panels (in Figure 8.7 and Figure 8.8) show the autocorrelation decreasing with lag suggestive that the posterior results would be accurate, and this is numerically portrayed by the effective sample sizes (ESS) (Kass et al., 1998) which are relatively high (at 1147.8 and 965.4 respectively.

The lower left panel in Figure 8.7 shows the shrinkage factor[99] which rapidly depreciates to a value close to 1.0. A value less than 1.1 is indicative that the between-chain (i.e. across chains) variance is similar to the within-chain variance, in-turn indicative of convergence of the solution (Kruschke, 2014). This is the case because if the variability within the chain is similar to that across chains, then the multiple chains would be considered of similar variability and as such stable and representative of the posterior estimates of the target distribution.

The density plot of the parameter values in the lower-right panels of Figure 8.7 and Figure 8.8, shows that the three chains produce overlapping densities, indicative of the representativeness of the results. Further the limits of the 95% highest density interval (HDI) (see Chapter 7, Section 7.4.1) for the density plots in Figure 8.7 and Figure 8.8 are respectively similar and this in turn is indicative of convergence of solution. Further, the numerical Monte Carlo Standard Error (MCSE) estimate for the variance in the chains indicates low values with MCSE 0.0405 and 0.0412 as shown Figure 8.7 and Figure 8.8 respectively. The high ESS couple with the low MCSE values are indicative that the MCMC designed for estimating the aggregate volume of passengers associated with Kirkstall Forge and Apperley Bridge stations are efficient. Note however, that the horizontal scale of the parameter distribution has been adjusted to reflect accurate volumes. The original scale in the data values were normalised by a factor deduced from a sensitivity analysis. This enabled the chains converge efficiently, and so the horizontal scale of the final parameter density distribution was re-adjusted to counter the original normalisation.

The robustness of the results shown in Figure 8.7 and Figure 8.8 are validated by estimating passenger flows at known locations in close proximity to the new stations. Shipley and Bingley train stations were chosen for this purpose and these served as reference points that enabled further balancing and fine-tuning of the model. Further, two additional reference railway stations (Leeds and Wakefield) were chosen to calibrate the model.

[99] The shrinkage factor is also called the potential scale reduction factor, and is referred to as the Brooks-Gelman-Rubin statistic. The intuition behind the statistic is that if the chains are stationary and by implication have settled into a representative sampling, the variance within the chain would be the same as across chains. In such circumstance, the ratio of the variances would approach 1.0.

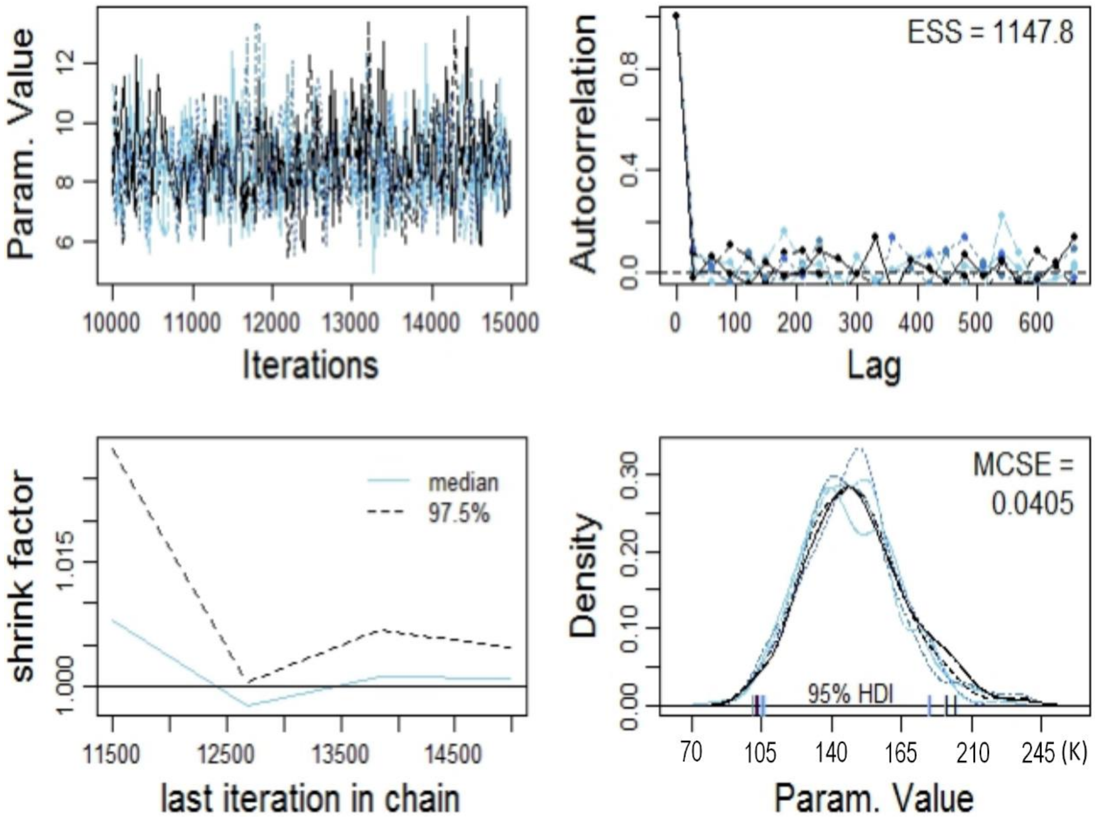


Figure 8.7 Aggregate passenger volumes at Kirkstall Forge.

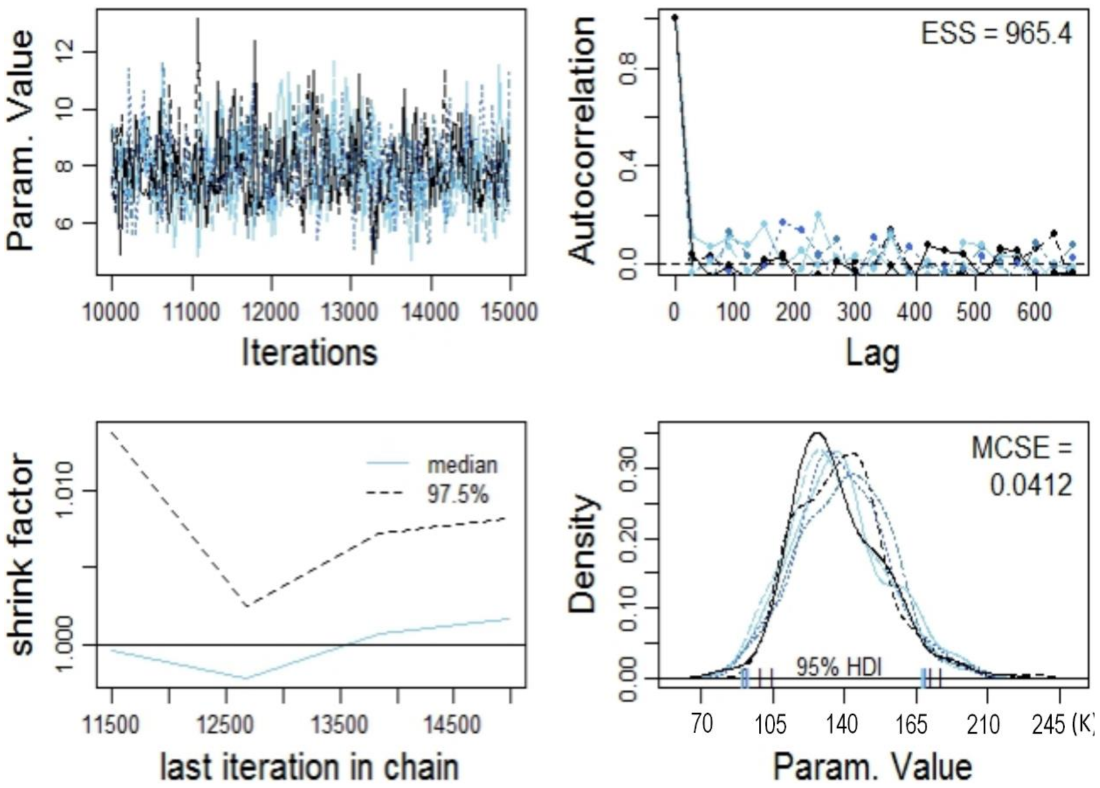


Figure 8.8 Aggregate passenger volumes at Apperley Bridge.

8.4.2.2 Disaggregate OD flows to Kirkstall Forge

The disaggregate origin-to-destination (OD) flows associated with the new Kirkstall Forge railway station were estimated using the model in Figure 8.5. This Kriging model is hierarchical in the origin and destination covariates. The disaggregate volumes of passengers are achieved by including the origin and destination stations using the nested indexing facility of the BUGS Bayesian modelling software. The details of the model in Figure 8.5 is included in Appendix E.2 alongside the R-script model wrapper. The results are shown in the 3D OD matrix shown in Figure 8.9 and Figure 8.10.

These disaggregate results facilitate ascertaining the inflows and outflows between the new Kirkstall station and all other stations in the study area. The disaggregate mobility is facilitated by including origin and destination hierarchies (the nested indexing) within the kriging model. This enables the disaggregation of conventional point phenomena into its constituents associated with particular origins and destinations. These modelling methods highlight the flexibility of the Bayesian framework in capturing the complexities of mobility. A further disaggregation to identify particular passenger types within the estimated volumes for the new rail station, can be achieved by including hierarchies of particular individual-level attributes.

The plot in Figure 8.9 shows the OD flows in the West Yorkshire study area prior to the introduction of Kirkstall Forge station (and Apperley Bridge station). The plot in Figure 8.10 shows the estimated OD flows in the West Yorkshire study area sequel to the introduction of Kirkstall Forge station (and Apperley Bridge station). Flows associated with Kirkstall Forge station are shown at the extremes of the horizontal and vertical axes. The higher flows are associated with the extreme red colour, while the lower flows are light yellow and then green for the lowest flows. The white boxes represent low flows which were zero due to numerical rounding (see below).

It is noteworthy to point out that the OD data used to estimate the demand at the new Kirkstall Forge station was created by aggregating the count of passengers within the different variable-categories of passengers on particular OD flows. This count was further normalised[100] and rounded to zero decimal places (as they ought to be integer counts for the Poisson hierarchical Bayesian model).

[100] The passenger count had to be normalised by dividing by a large factor (1000) to enable the Bayesian regression model to execute.

Apart from the normalization of the count variable, the travel distance and geography scale had to be similarly normalised by dividing the appropriate variables by 1000. Otherwise, it was not possible to execute the Bayesian model for the combinations of priors and initial conditions explored. The need for normalisation was a result of the sensitivity of the solution to the extreme counts, thereby causing a failure of the algorithm. It was thought that this large variation in counts might have affected an ability to specify appropriate parameters for the prior distribution. It was not possible within the timeframe of the project to fully explore this issue. A compromise was to normalize the count data, travel distance, and geography scale, then the Bayesian model was executes and the results appropriately re-scaled.

The normalisation did not affect the structural of the Bayesian model, however, a division by a factor of 1000, and then rounding off to zero decimal places would have resulted in data loss. This inadvertently also reduces the accuracy and precision of the results, and it was not possible to preserve the total passenger flows. As a result of this, the flow trends estimated were reported as shown in the heat map OD plot in Figure 8.9 and Figure 8.10. The validation of the results were established by comparing the flows estimated using the Bayesian Kriging model with MOIRA flows estimated since the introduction of the new Kirkstall Forge station. As a result of the normalisation and rounding to zero decimals, the flows predicted for those stations with low passenger counts were zero and not accurately indicated on the OD heat map (see Figure 8.9 and Figure 8.10). MOIRA results for stations with largest flows are 10.7m for Leeds (LDS) and 8.6m for Huddersfield (HUD), while Kirkstall Forge (KLF) was 108k. The estimates from the Bayesian kriging model were 3.2m, 2.3m, and 94K respectively.

As the MOIRA flows tended to be about 2.5 times the simulated estimates, it is expected that the 94k estimated from the Bayesian Kriging would actually amount to about 250k flows if multi-modal tickets and flows beyond West Yorkshire were included. The Kirkstall Forge station opened in 2016, and the MOIRA estimate for entry and exit flows once the station actually opened increased from 95k during 2016/17 to 150k during 2017/18. The anticipation is that based on the Bayesian Kriging results from this research, the Kirkstall Forge flow volumes will stabilize at about 250k per annum in the coming years if the conditions on the rail network remain stationary.

The software code used to estimate the impact of the new station at Kirkstall Forge is described briefly in Appendix E.2, and the R-script code used to generate the results are included in the CD accompanying the thesis.

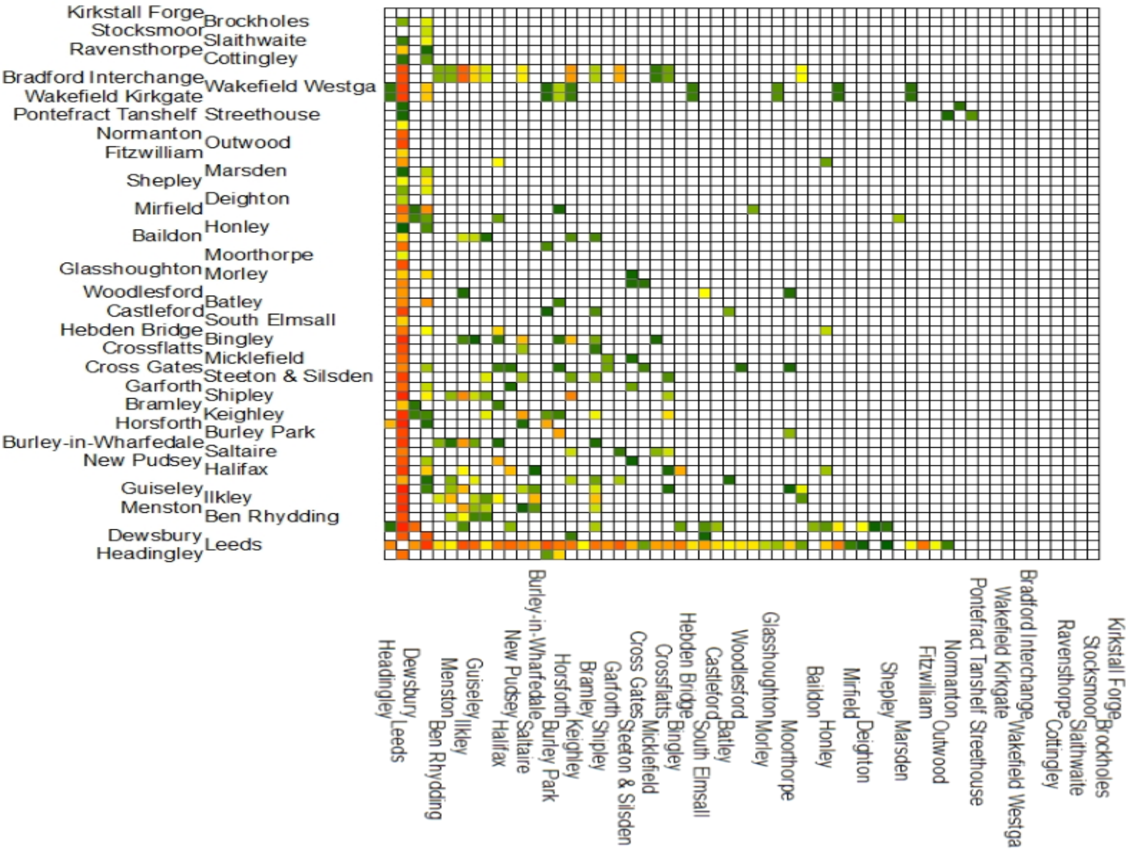


Figure 8.9 3D OD plot without Kirkstall Forge flows.

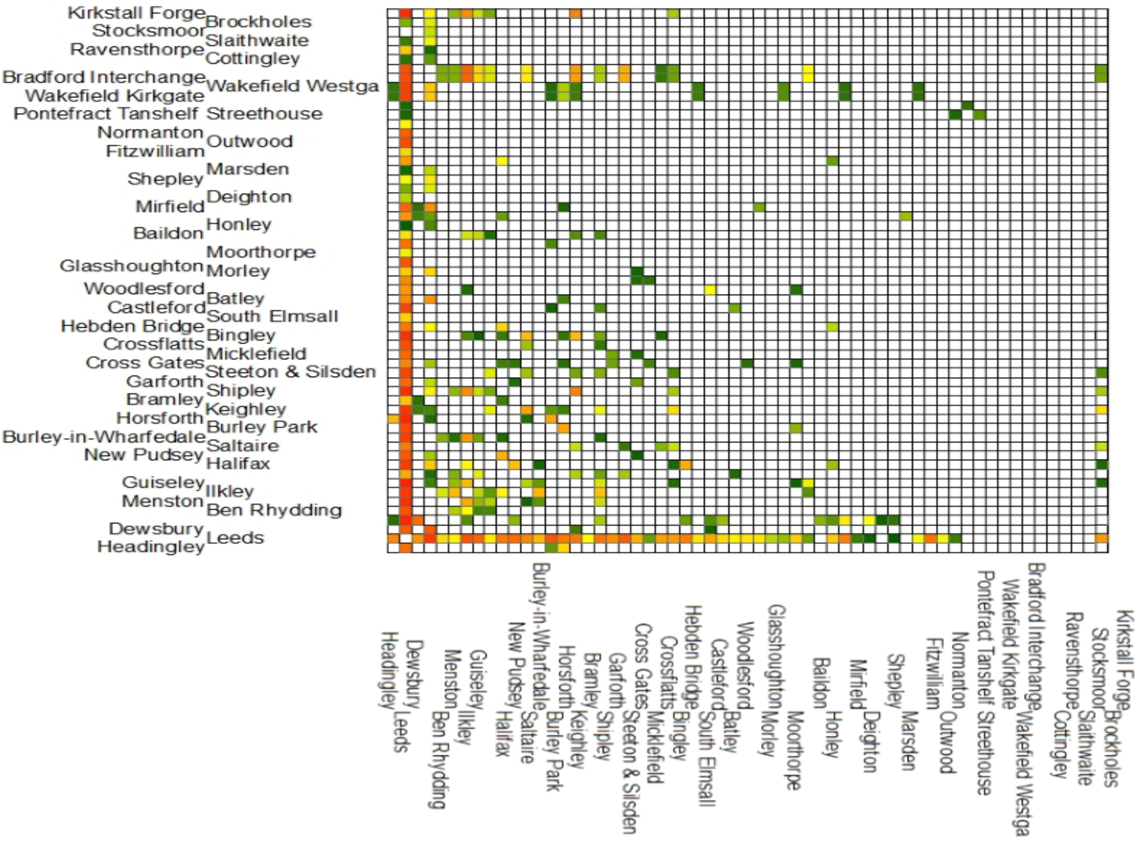


Figure 8.10 3D OD plot with Kirkstall Forge flows included.

8.5 Remarks

The results from the case study of rail-heading in West Yorkshire, and the impact of a new station at Kirkstall Forge illustrates the strengths of the Bayesian framework in constructing a parsimonious model that captures the nascent features of mobility phenomena on the railways. In the case of rail-heading, Bayesian analysis provided the ability to explore the contribution, to the mechanism of rail-heading of geographic areal attributes, and the contribution of heterogeneity in latent attributes. In the case of the new railway station at Kirkstall Forge, an ability to disaggregate mobility into hierarchies enabled the identification of OD flows between the new stations and all other stations on the network. In the conventional kriging framework, only aggregate volumes of flows would be feasible as no way exists of disaggregating flows from the same geocode.

When Bayesian analysis is applied to complex realistic data like the railways data, knowledge of the data generation process is paramount in constructing an objective mobility model that would converge efficiently, and is accurate, stable and representative of the posterior distribution. Apart from this, knowledge of the data aids in the specification of initial values, and priors on the model parameters. Sometimes, as has been the case in this research, simple frequentist models can give indications of the parameter values, and guide the choice of initial values and priors. These practical steps have aided in model development, but to date no universal Bayesian model development process has guaranteed convergence within specific time constraints.

The Bayesian mobility analysis presented in this chapter is not aimed at being exhaustive as focus was on simulating and imputing a micro-level representative population. However, aspects of the modelling have been presented to illustrate some of the strengths of Bayesian models, to enable reproduction of the results, and to potentially facilitate further investigation. Apart from the strengths of the Bayesian approaches, the disadvantages are most notably the high technical skills required to specify the models, the priors, the initial conditions, and to run the models on WinBUGS, OpenBUGS, JAGS, and Stan. These tend to require advanced software skills and statistical knowledge, and a firm intuition about the solutions tenable, with no guaranteed convergence. The high computation load in terms of RAM requirement for large data tables and arrays, and the computational time (often taking up to 10 days on a supercomputer!) all add to the challenges in adopting Bayesian modelling.

In the case of flows emanating and terminating within West Yorkshire investigated, the phenomena of rail-heading observed in the vicinity of Brighouse and the Bradford stations (see Figure 8.6) are driven by passengers accessing the train service at Wakefield to minimize the number of stops and to an extent travel time. A more detailed behavioural research is needed to further understand idiosyncratic details of why passengers exhibit rail-heading behaviour. This research has demonstrated that the phenomena exists, and can have a structural Bayesian model fitted to it.

The maps in Figure 8.11 and Figure 8.12 support the mechanism of rail-heading identified by Bayesian analysis. The heat maps illustrate the zonal concentration of train stations in Figure 8.11, and the frequency of train services in Figure 8.12. The darker regions are respectively indicative of higher density and frequency of trains at the different zones. Figure 8.11 and Figure 8.12 combine to give an indication of zonal access to train service in the West Yorkshire study area.

As seen in Figure 8.10, the density of rail stations is low between the vicinity Brighouse and the Bradford stations. However, there are commensurately higher densities between Wakefield or Brighouse stations, and Leeds. This, coupled with the high volume of train services in the vicinity of these stations (Brighouse and Wakefield), might have prompted the choice of access for trips to Leeds. Further, the Brighouse station is in Rail Zone 4 (see Figure 8.1), whilst Wakefield is in Rail Zone 3, making Wakefield ultimately the cheaper choice for trips to Leeds. An investigation of the train timetables and fares (Google, 2017, National-Rail, 2018, WYCA, 2018) in the period around 4.30 pm when the rail-heading is prevalent confirms that the Brighouse station is served less frequently, this coupled with the higher fares perhaps encourage passengers to seek alternative modes to access stations that are better served during that passengers choice of travel time. These insight are useful in decisions about periods where it would be of beneficial to adjust the train timetables (typically done during the fare rounds).

These analyses discussed above are not exhaustive but preliminary and indicative. Due to the challenges in developing current state-of-the-art Bayesian models, a considerable amount of time is spent in the development processes. As such currently Bayesian approaches are sold as a method of providing confidence in existing industry assumptions, and also as a method that challenges existing industry models.

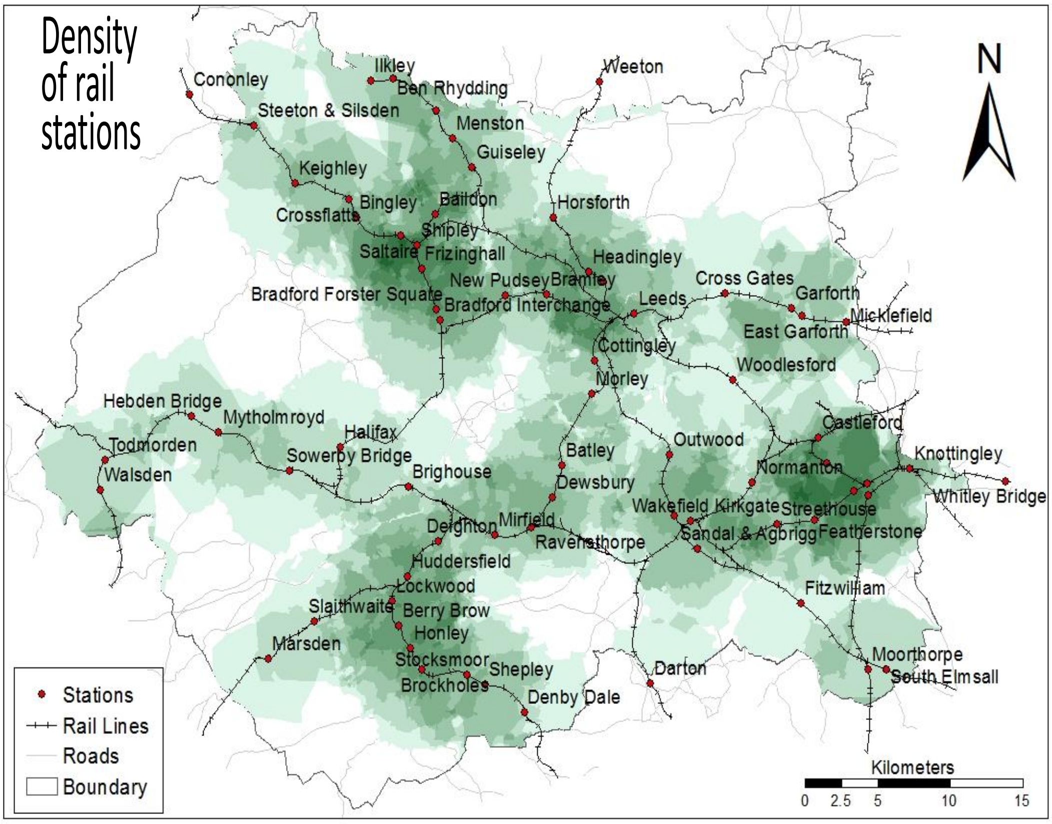


Figure 8.11 Density of train stations in West Yorkshire.

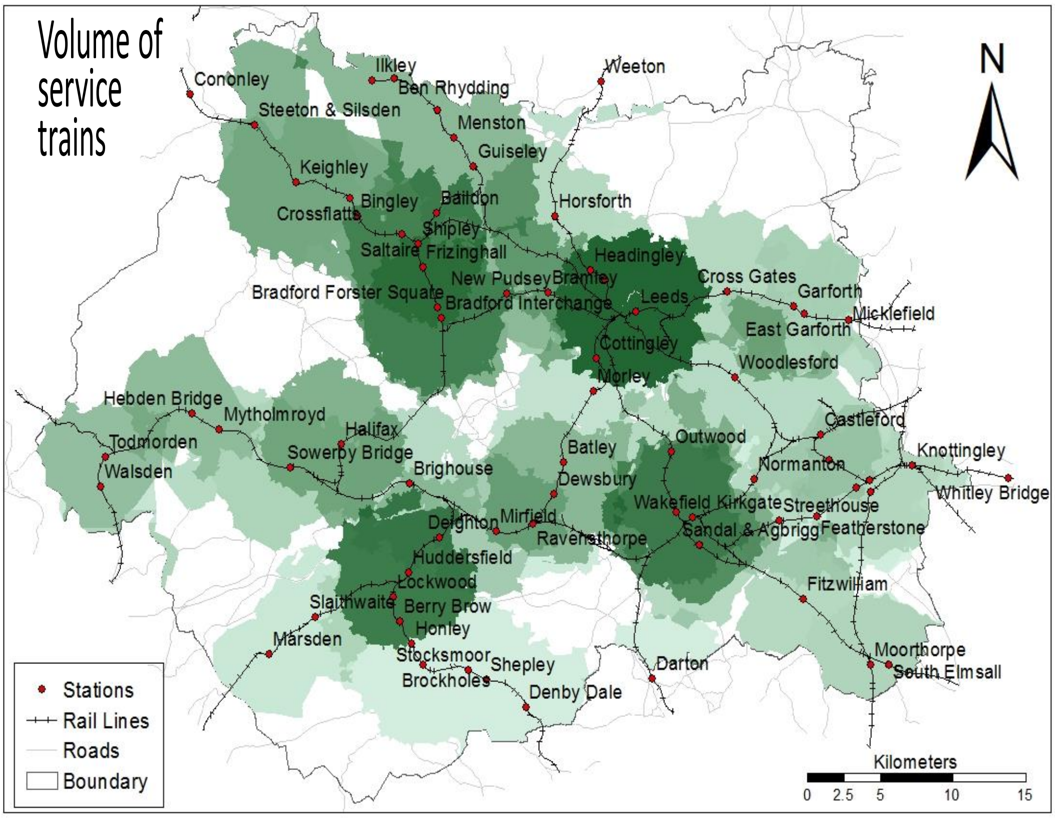


Figure 8.12 Volumes of train service in West Yorkshire.

The impact of a new station a Kirkstall Forge has been identified using a hierarchical Bayesian model developed within the geo-statistics kriging framework. The hierarchy within the model enables the disaggregation of the point mobility into constituent parts forming the volumes of passenger interacting with other train stations within the West Yorkshire study area. Project exigencies have meant that the mobility model created included the main effects origin and destination variables, and the distance variable as is typically used in traditional spatial interaction models. Within Bayesian models, it is possible to incorporate additional variables to assess their influence, but these were not implemented in this research due to challenges in ensuring convergence of the MCMC model sampler within the confines of time on the project. Efficient convergence of Bayesian models has been a particular handicap to an otherwise robust analysis framework.

MOIRA data recently made available to the project has indicated that the volumes deduced for the new Kirkstall Forge station are of the order of aggregate magnitudes derived from MOIRA (see Section 8.4.2.2). This serves as an external validation of the hierarchical Kriging model developed to investigate the impact of a new railway station at Kirkstall Forge. The difference in magnitudes is attributable to the exclusion of PTE tickets and the limitation of the flows to those that emanate and terminate in West Yorkshire.

The Bayesian modelling strategies are gaining increasing promise because of the range of flexible features that enable an investigation of a range of case studies. The uncertainty associated with the results can also be assessed from the distribution of the parameter values deduced. The Bayesian framework includes these constructs and frameworks that make them suited to the analysis of novel rich micro-level consumer datasets. It would for instance be easy to implement so called evidence synthesis where there are many disparate datasets available to construct models (Gelman et al., 2014a, Kruschke, 2014, Lunn et al., 2012). These features make Bayesian models pertinent in the analysis of consumer data.

In this chapter, a number of arguments have been put forward about why passengers partake in rail-heading. These arguments are by no means exhaustive. It has not been possible in this research to understand fully the mechanism of rail-heading. Further studies would need to be conducted to get to the underlying choice drivers of rail-heading. The types of social surveys involved are beyond what is achievable in the time constraints of this research.

Chapter 9
DISCUSSIONS AND CONCLUSIONS

For completeness the ethical issues related to large consumer data (so called big data) are discussed. Ethical issues are important considerations to be made in developing a framework for the use of disparate consumer datasets. The ethical issues highlight the regulatory and legal frameworks in place to foster diligence in handling consumer datasets. Knowledge of these ethical concerns will eschew safety and security of datasets. This will help in the sustainable availability of data in the future for research. This chapter then presents the main conclusions of the research providing the final word on the value of the analysis conducted in the research as presented in the thesis. The potential impact of the research is presented in the conclusion, along with a tie in with the aims and objectives of the research, in such a way as to reflect the abstract of the thesis. Then the summary of the thesis is presented, highlighting the relevant findings from the research. A synopsis briefly summarises the details presented in the chapters of the thesis. The summary highlights and presents a general survey of the pertinent issues, and a brief discussion reviews some of the concepts behind the powerful methodologies developed, drawing special attention to how each of the concerted methods complement other methods adopted.

Work anticipated forming a future research agenda is presented, highlighting areas of development to tie in with current related research activity. The potential availability of additional consumer datasets presents an opportunity to extend the investigation beyond the boundaries of the current study area. This in turn facilitates an analysis of additional boundary mobility phenomena. The stochastic MCMC based population synthesis methods are areas of future work to ensure their suitability for application to skewed consumer data. The Bayesian framework has potential for application to the analysis of a wide range of complex mobility phenomena. Being based on the ArcGIS platform, the GIS-GTFS network simulation developed in this research have potential for use as a tool for interventions assessment on the rail network, and as an educational tool.

This chapter underscores the limitations of the work conducted during the research. The future application of the methodologies developed, and the impacts of the research on the railway industry are discussed.

9.1 Ethical issues

Ethical issues are of paramount importance in the use of consumer big data in research on urban mobility and movement patterns. The UK government had a Data Protection Act 1998 which established laws for public bodies to encourage openness in data access, uphold information rights in the public interest, while promoting privacy rights for individuals (ICO, 2015). The Freedom of Information law (OGL, 2015) defines the public's rights and procedures for access to information held by public bodies. The Human Rights Act 1998, especially Articles 8 and 14 prescribes an individual's inalienable right to respect for private and family life ensuring that private data on individuals have to be dealt with surreptitiously so as not to contravene the privacy laws or violate particular ethical discrimination provisions.

The Statistics and Registration Services Act was introduced in 2007, establishing the UK Statistics Authority (UKSA) supported by the Office of National Statistics (ONS), to establish regulations guiding the sharing and confidentiality of personal information. Datasets within the UKSA and ONS form the most complete sets of data within the UK, and access to such data for this research requires a subscription to strict guidelines laid by the ONS. The UK Environmental Information Regulations of 2004 was derived from European law (ICO, 2015) and covers spatial information, information developed through the EC's INSPIRE initiative (INSPIRE, 2015), environmental information including CCTV recordings, and other softcopies held by public authorities including universities, as well as data from vehicle automatic number plate recognition (ANPR) devices. The Environmental Information Regulations (EIR) prescribe directives to relevant authorities for storing and providing access to such datasets, and the CDRC which is the parent centre for the research project reported in this thesis abides by the EIR prescription.

By virtue of the above UK laws governing data (ARCHIVE, 2011), consumer data of a sensitive and confidential nature related to the urban mobility project can safely, ethically and legally be shared if the Data Protection Principles (JUSTICE, 2008) are adhered to. The principle consists of important aspects considered jointly during the processes of data procurement and use. It requires that data be procured with the consent, within the expectations of use (for public, government or commercial benefit) and with the forbearance of the subject of the data. The data procurement and processing should be seen as satisfactory and fair to the subject of the data. The purpose for which the data was collected and processed should be lawful and properly specified such that the data is adequate, relevant and not self-indulgent. The personal data requires being accurate, properly labelled, dated and kept secure.

In the UK, government regulations preclude organisations from keeping personal data for longer than necessary and requires anonymising the data to protect the identity of the data subjects in accordance with individuals' rights (ARCHIVE, 2011). The data should not be transferred to a non-European Economic Area (EEA) country, and within the EEA, access to the data should be controlled and managed according to regulatory standards. An application to use the secure 2011 UK Census data is limited to September 2015 to 2016 with strict guidelines on extending this period.

Big data for urban mobility research often come from secondary individual customer information (like ticketing data), and in many instances as discussed above, there could be ethical implications of not adhering to strict rules concerning security and safety of the data. Big data is novel, and particularly attractive, and the moral principles that govern access to such information and consensual use of such data is not firmly established. This calls for a brief review of ethical guidelines developed by governments, public (including research institutions) and private institutions that intend using novel big data (Rayport, 2011). In many instances of concern, the private authority already has access to private individual information by virtue of the services they provide to the individual and the lure of identifying the mind-set of the customer by studying behavioural data often may overwhelm these institutions into unwholesome practices.

Similarly, external organisations in a bid to secure commercial advantage have sought to aggressively accessing data of the competitors, without being fully aware of the legal, safety and ethical consequences. A current incident which springs to mind concerns Cambridge Analytica (Cadwalladr and Graham-Harrison, 2018), whereby Facebook (Ellison et al., 2007) individual data was purportedly unduly acquired and analysed to develop models of voting patterns and the electoral process. This has led to the call for institutions to employ data officers (as with the CDRC), and to introduce additional consumer privacy laws, online tracking laws, hacking laws, rights to be forgotten laws, and the EU Article 29 guidelines for electronic surveillance for national security and inter-state exchange of data. The UK along with the EU are currently (as at the time of publication) introducing the new General Data Protection Regulation (GDPR) (Kuner, 2012, Mantelero, 2013), a framework for personal data protection laws. The aim is to harmonise data privacy laws, and supersede many of the outdated personal data laws across the EU. The GDPR framework came into force on May 25, 2018, and in the UK a new Data Protection Act 2018 has also come into force (UK-Legislation, 2018). Figure 9.1 is a summary of the salient aspects of the GDPR, highlighting its prominence in issues related to consumer data safety and security in the EU.

Public and private institutions like the ESRC funded CDRC involved in using big consumer data about people for research have significant ethical and sometimes legal ramifications concerning discrimination, sexism, racial bias and bigotry. This is especially the case as in big data analytics, geographic modelling, and causality simulation whereby different dataset may have behavioural relationships to particular sections of society. The University of Leeds has developed academic integrity standards as a bedrock for ethical research which extends to the use and sharing of data, and the university now has an established research ethical review process (UoL, 2015). The Consumer Data Research Centre (CDRC) subscribes to the ethical standards of the university.

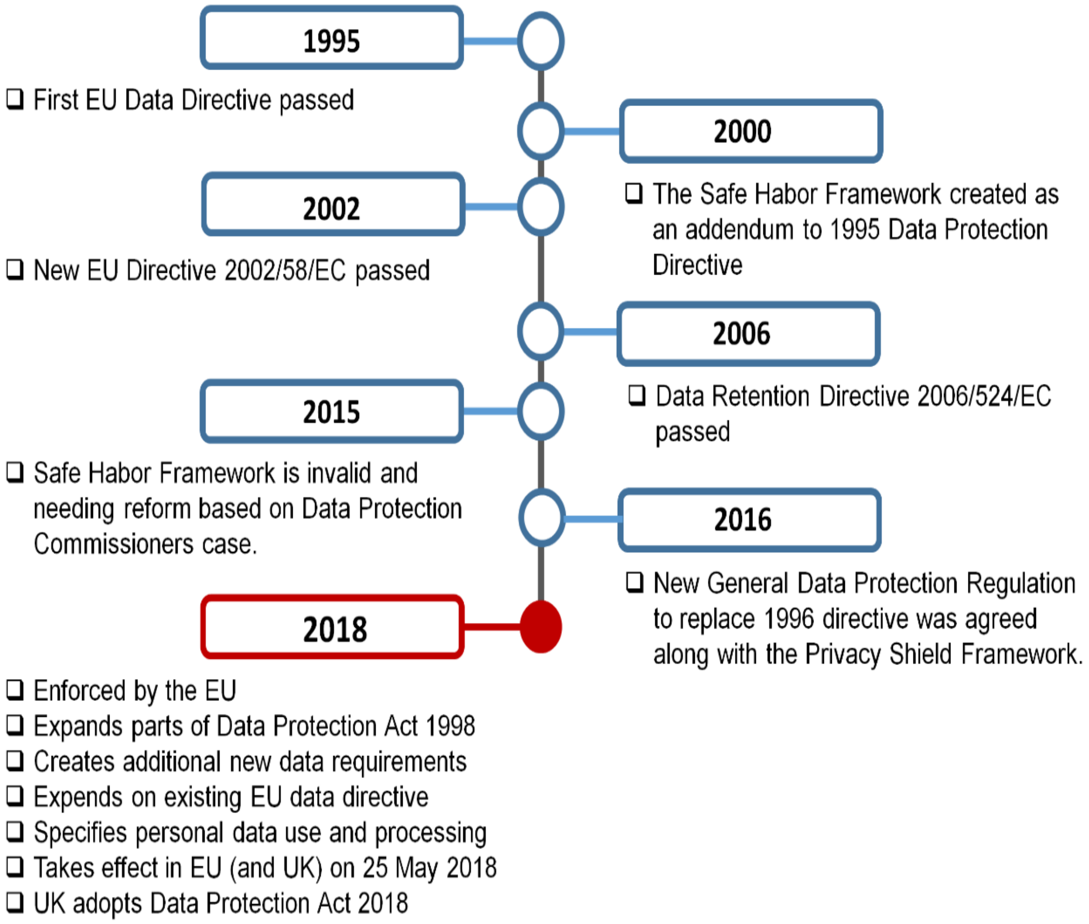

1995
❑ First EU Data Directive passed
2000
❑ The Safe Habor Framework created as an addendum to 1995 Data Protection Directive
2002
❑ New EU Directive 2002/58/EC passed
2006
❑ Data Retention Directive 2006/524/EC passed
2015
❑ Safe Habor Framework is invalid and needing reform based on Data Protection Commissioners case.
2016
❑ New General Data Protection Regulation to replace 1996 directive was agreed along with the Privacy Shield Framework.
2018
❑ Enforced by the EU
❑ Expands parts of Data Protection Act 1998
❑ Creates additional new data requirements
❑ Expends on existing EU data directive
❑ Specifies personal data use and processing
❑ Takes effect in EU (and UK) on 25 May 2018
❑ UK adopts Data Protection Act 2018

Figure 9.1 Development of General Data Protection Regulation.

9.2 Research conclusions

This research set out to develop a framework for use in harnessing large consumer data (so called big data) available from the railways, for urban mobility and movement patterns analysis. The objective was to harness the consumer data to create a requisite dataset, rich enough in exogenous and endogenous attributes, that they can be deployed to the range of trip, activity, and multi-agent based transport modelling concepts. Traditional consumer data available from the railways is the passenger ticketing data, typically at individual-levels. Harnessing this dataset yields a representative, attribute-rich synthetic micro-level population of railway passengers. Studies on such micro-level data enable the identification of individual-level mobility and movement patterns on the railways to enable better forecasts of passenger demand. Any missing passenger behavioural attributes can further be imputed by adopting robust Bayesian methods based on statistical best practices, for imputing missing values. Bayesian models apart from their use for imputation, represents new analytical method for spatial analysis consumer ticketing data, revealing the mechanism of a range of mobility phenomena on the railways network and the wider environment.

This research explores opportunities of reviewing modelling tools, using big consumer ticketing data available within the UK railways, to develop new methods for understanding, simulating and predicting individual movement patterns within the railways, considering frameworks relating population mobility to demographic characteristics, relative location, time of the day, and other likely drivers of behaviour and interaction. A set of three complementary concerted methodologies were developed for these: spatial microsimulation of skewed data, GIS-GTFS network simulation incorporating real-time transit scheduled, and Bayesian imputation and modelling, incorporating special modelling construct to better reflect the complexities of mobility. The tool developed facilitates meeting the challenges of efficiently operating the railways while ensuring economic, social and environmental sustainability. These innovative methodologies are requisite to the analysis of the multi-variate inter-relationships between the facets of the network infrastructure, passengers and surrounding habitats, and to visualise these railway network mobility interactions. The rich micro-level synthetic population can potentially have a significant impact on the quality of inputs to strategic, tactical and operational rail-sector planning models. The GIS-GTFS model is potentially useful to network operators for the management, maintenance and interventions assessment on the railway network.

A powerful set of complementary concerted methodologies have been developed in this thesis to harness consumer (ticketing) data available from the UK railways. The first of the methodologies is spatial microsimulation which is used to combine LENNON ticketing, NRTS and Census interaction data, to yield a micro-population inheriting attributes of the combined datasets. A novel spatial microsimulation methodology suitable for typically skewed consumer interaction data is developed. The second methodology created is a logistical GIS network which incorporates information on transit schedules within the publicly available general feed transit schedule (GTFS). The micro-population created from the first methodology is used as input to the logistical GIS-GTFS network, to reconcile the simulated population with the context of passenger mobility within the rail network. The resulting micro-population consists of railway passengers embedded in the wider population, and this forms a dataset rich in endogenous and exogenous attributes, useful for subsequent analysis. The third methodology, the Bayesian framework is used as the imputation model to identify missing daily trip rates with season tickets, flows to group station and other idiosyncratic behaviour of individual passengers. Bayesian models are further developed to investigate a range of case studies of spatial analysis of urban mobility and movement patterns in the West Yorkshire study area.

The framework presented in this thesis for the use of big data in urban mobility and movement patterns analysis, complements and provides an alternative to the traditional four-stage paradigm for transport modelling. The availability of novel data from consumers of digital services in many instances can supersede the need to formulate the generation of initial trips in transport models. Consumer datasets like the LENNON tickets are often large and comprehensively include all the trips, precluding the need to construct or generate a volume of trips. Further, knowing the origins and destinations of every ticket sold, precludes the need for the second classic transport modelling stage of trip distribution. For the consumer ticketing data, the trip distribution is already captured during issuing of tickets from the over 8000 ticket sales facilities online, within station and at retail outlets. The Bayesian framework provides opportunity to impute and analyse mobility in one step. It is advantageous to identify missing values in the same context from which the data was generated. Within the Bayesian regression, the imputed data values can be tailored to values informed by experience (in industry applications), by innovatively using simple and effective constructs like censoring, truncation and mutual exclusivity of parameter values. These enable the creation of more realistic models (Lunn et al., 2012).

9.3 Summary of thesis achievements

9.3.1 Spatial microsimulation

The detailed heterogeneity in large consumer datasets can be better harnessed by integrating them with other relevant datasets from measured stated preference surveys which are designed to be random samples representative of the population. In Chapters 4 and 5 of this thesis, both deterministic and stochastic spatial microsimulation strategies are discussed for combining various datasets, to create a representative micro-mobility population of railway passengers embedded in the wider population within a geographic region (West Yorkshire, UK). A deterministic methodology particularly suited to adjust skewed rail-mobility consumer big data is developed, combining information on all rail tickets sold in the UK, with the 2011 Census commute to work data and a National Rail Travel Survey (NRTS), yielding a representative micro-level population. The simulated population created includes weights which represent the probability density for rail commuters, such that a sample of the synthetic population according to the density yields representative rail passengers.

In transport modelling exercises, demand estimation generates a synthetic population and is a core component of the modelling process, critical to the accuracy of any subsequent transport simulation. Our research makes fundamental contributions to the demand generation process by creating an attribute-rich micro-level representative population. In particular, microsimulation papers in the literature (Guo and Bhat, 2007, Müller, 2011) highlight the challenges in correctly simulating a population due to the so called ‘zero-cell-value’ problem and the difficulty in simultaneously controlling for more than one level of constraint (say individual and household within a zone, i.e. a 2-level constraint hierarchy). This thesis proposes and develops for the first time a methodology which as well as resolves the peculiarities of combining typical consumer data with other measured surveys, resolves the demand population issues. The highlights of the spatial microsimulation methodology are as follows:

- Development of a methodology for harnessing typically skewed large consumer big data available from the railways.
- Investigates, for the first time, the behaviour of deterministic and stochastic microsimulation strategies when, as is typically with novel big data scenarios, the seed data are skewed relative to the target distribution. Our research has identified the suitability of the multi-

dimensional IPF (m-IPF) strategy. We have highlighted the mechanism through which the m-IPF strategy converges to a stationary distribution of a desired target population when the seed is skewed. We point out why the stochastic MCMC based methods converge to the distribution of the seed data instead, and suggestions are made about strategies for developing the MCMC based stochastic methods for application to skewed seed samples.

- Validation of deterministic and stochastic spatial microsimulation methodologies, and explanation of the mechanisms that results in preference for the deterministic (IPF) procedures (over stochastic ones) in applications for novel consumer data.
- Underscores that in a typical big data scenario, the weights derived from spatial microsimulation are reflective of how representative individuals are of an origin-destination flow, and importantly also are reflective of the levels of skew (bias) in the big data (used as seed). In the literature on IPF strategies (Guo and Bhat, 2007, Müller and Axhausen, 2010), it has traditionally been assumed that the IPF weights are only an indication of how representative an individual is of a zone of assignment. This distinction gives a better insight into the spatial microsimulation mechanism.
- Simulation of a demand population that better reflects mobility interaction by applying simultaneous constraints to origin and destination attributes in the IPF procedure (a so-called 3-level constraint hierarchy). Simultaneously constraining for origin and destination creates a conceptually consistent demand population by including the mobility interaction in the spatial microsimulation process. To the best of our knowledge, applying this novel strategy to big data would be the first time such method is presented in the literature.
- Ensures that the aggregate and seed data always have matching seeded variable-categories and as such reveals any potential zero-cell value. Further, the methodology developed in this research simultaneous applies a 3-level hierarchy, enabling an exploitation of the full joint distribution of the target dataset (as opposed to just the marginal).The hierarchy is made up of an individual, as well as their origin and destination, forming the 3-level simultaneous constraint. This methodology is as such novel in extending current 2-level hierarchies in literature (Müller and Axhausen 2012), and also in applying this to novel big consumer data available in the railways.

9.3.2 GIS-GTFS network simulation

Apart from not having a full repertory of passenger attributes, the ticketing consumer data available from the railways does not have information about the context in which the ticket was used (such as the level of crowding in the train when the ticket was used, or the waiting times the passenger endured). This lack of context presents challenges in the planning and management of railways, and in causal inference on the drivers of mobility choice. The second of the three concerted complementary methodologies developed in this thesis, utilises the synthetic population from spatial microsimulation as input to a GIS model. In Chapter 6, the GIS network model is developed. The GIS network model is logistically constrained by transit schedules now made publicly available as GTFS information. The new GIS-GTFS network model accentuates the application of new and improved analysis techniques to transfer big consumer ticketing data into useful decision support information for the railway sector. This is achieved by creating a dataset containing contextual information on passenger mobility.

The research presented in this thesis contributes to the state-of-the-art by the creation of a network simulation model based on consumer big data and GTFS transit schedule information. To the best of our knowledge, no application exists that has exploited railways ticket information in conjunction with the GTFS information for network simulation to identify the context of passenger mobility using actual railway ticket information. The aptness of the contribution is highlighted when it is considered that, even established transport simulation tools like MATSim have only recently (Poletti et al., 2017) started to incorporate GTFS information within the modelling framework, making the contribution presented in the thesis quite timely.

The GIS-GTFS methodology presented develops a range of novel methodologies and data analysis techniques with results highlighted below:

- Development of a network simulation model based on consumer ticketing big data that are reconciled with novel transit schedules.
- Development of GIS-GTFS network analysis to create a set of rich endogenous attributes, enabling the assignment of passengers to the network traffic, and avoiding the complexities of utility optimisation for traffic assignment.
- The creation of a rich demand data implies no need for the utility traffic assignment adopted in transport modelling, potentially altering traditional modelling genres

- The GIS-GTFS model enables visualisation and animation of mobility of both passengers and transit trains on the railway network.
- Representative attribute-rich simulated populations (from tickets, Census, NRTS surveys) are best suited for subsequent use in mobility regression analysis.
- The GIS-GTFS network model is useful for management, maintenance and interventions assessment on the railways. Within the traditional GIS paradigm, such models are invaluable for use in inferring 'what if' scenarios on the network.

In addition, the GIS-GTFS network created demonstrates the mechanism of mobility evolution on the rail network. This presents in a practice oriented way the accurate assignment to the railway network traffic, the rich synthetic population which contains information on passenger arrival time at train station, first train, intermediate stops, final destination, ticket restriction, access and egress mode and distances, as well as a plethora of addition information. This precludes the need to conduct the complicated utility optimizing traffic assignment. Learning, feedback and adaptation are traditional transport modelling tools used for assigning passenger journeys (i.e. flows) to the traffic network.

MATSim and EMME (Consultants, 1999) again referenced in this instance as a state-of-the-art tool, have core expertise in traffic assignment through feedback and adaptation of the flows deduced from spatial microsimulation. In both methods the passenger effectively re-assesses their strategy for arriving at an origin and destination (OD), and in some cases changes the OD. In MATSim each passenger has a set of activity schedules assigned and these are stochastically selected and perturbed until a predefined marginal utility objective function converges to a stationary equilibrium (Gao et al., 2010, Horni et al., 2016). The EMME assignment is conversely a deterministic procedure which iteratively arrives at an optimal solution satisfying a non-linear flow-travel time objective function (Barthélemy, 2011, Spiess and Florian, 1989). The strategy presented in this thesis introduces a different perspective to transport simulation by creating a demand datasets rich in exogenous and endogenous attributes such as to avoid the complexities of the utility optimisation stage for traffic assignment. As such, our strategy exploits the rich attributes in the big-data-based synthetic population in deciding each passenger's assignment, better reflecting the congestion points within the network.

9.3.3 Bayesian modelling

In Chapter 7, the Bayesian modelling framework is presented, and in Chapters 8, two spatial analysis case studies were investigated using the Bayesian models. The particular strength of Bayesian modelling as implemented in BUGS is its flexibility (Lunn et al., 2012), and an ability to create models that capture the range of complex phenomena. Bayesian modelling is attractive because of its flexibility in specifying constraints on values, and its hierarchical and interaction modelling capabilities.

Bayesian strategies explicitly ascertain the uncertainty associated with each inference, and are easily setup for posterior predictions to answer the many what-if situations that arise in case studies. In imputing journey factors for instance, industry experience indicates that daily journey rates on season tickets do not exceed twice, making it objective to truncate journey factors to two especially for lengthy journeys. Tickets of flexible periods of validity are usually used for at least the minimum validity period, making it sensible to censor the imputed validity periods accordingly. Whilst single and return tickets are restricted to use once and twice respectively, season tickets use on the other hand would normally fluctuate around the period of validity, making it appropriate to model journey factors as a finite mixture. Augmented flows were created to represent flows to group stations, in the proviso that these flows represent a single flow and as such add to one flow. The '*dsum*' Bayesian construct made simple provision for representing such a feature in the model. All these Bayesian modelling constructs enabled the creation of an objective model incorporating the typical traditional frequentist constructs like Poisson, Negative Binomial, and Bernoulli and Binomial regression.

The Bayesian concept of applying a large number of relevant variables in the imputation is premised on Rubin's work (Rubin, 1976), whereby the set of observed variables are sufficient to explain the missing values. To further enhance the application of this concept, the Bayesian framework enables the introduction of modelling constructs to ensure that the parameters and values deduced from the modelling exercise better reflects empirical evidence from the wealth of industry experience. This as such reflects a more robust modelling methodology.

The highlights of the Bayesian model include:

- Facilitate the identification of additional detail about the idiosyncratic behaviour of individual passengers, including: each passenger's daily journey rate, frequency of use of flexible tenure tickets within the period of validity, and individual passenger flows to particular stations in scenarios where stations are grouped together.

- Bayesian imputation methods complement logical rules based infilling methods currently applied in the rail industry as they exploit the variability and richness in observed data, (by utilizing a mobility model representing the data generation process) in the process of imputing missing values.
- Bayesian regression models are flexible in specifying constraints on values, and has established hierarchical and interaction modelling capabilities.
- Bayesian strategies explicitly ascertain the uncertainty associated with each inference, and are easily setup for posterior predictions to answer the many what-if situations that arise in case studies.
- The ‘*dsum*’ construct in Bayesian models enable augmentation of data and subsequent filtering out of the relevant dynamics from the augmented dataset.
- The Bayesian concept enables the application of a large number of covariates in the imputation process, and this is premised on Rubin’s work (Rubin, 1976), whereby the set of observed variables are sufficient to explain the missing values.
- The credible intervals deduced for the journey factors using the Bayesian methodology straddle the range of aggregate proxy values of journey factors adopted in industry. However, the Bayesian methodology disaggregates the journey factors to individual-levels. Thereby facilitating the construction of richer pictures of passenger behaviour, and the associated movement patterns on the railway network.
- The Bayesian framework enables the identification of the propensity of rail-heading within zones in the geographic study area. With the exclusion of boundary phenomena (whereby passengers travel to cross a travel zone to access cheaper fares), it is observed that within West Yorkshire passengers who have access to cars travel further to access a train station. The overriding decision for rail heading is to optimize travel cost, stops, and times on the rail network. The case study identified groups of passengers who travel from Postcode Sector nearer the Bradford stations to access the rail network further afield at the Wakefield stations en-route to Leeds station. It is observed that more regular high speed train services are available from Wakefield to Leeds station.

9.4Future research agenda

This research was part of an industry-academia joint venture. The academia side of activity was further intercollegiate and multi-disciplinary. The industry partners also had consultancies and data providers associated with the research activity. Whilst the industry interest leaned more to immediate practice-oriented applications results, the academia had to balance the needs of producing work commensurate with standard suitable for such research. Balancing the disparate demands of the multi-facetted project interest parties meant that a single work of this size and tenure would not cover all the potential avenues of research established and revealed during the course of the work. The following sections discuss briefly such avenues of research identified to be of potential interest for any future work, along with some implementation suggestions of such work.

9.4.1 Wider RSP ticket coverage and PTE tickets

There are a range of areas where the research presented would benefit from further work, primarily in conducting a more comprehensive study of the mechanism of rail-heading. Current rail-industry indications are that the rail-heading[101] phenomena is prevalent at rail zone boundaries, due to passengers travelling further to access the rail service within the West Yorkshire rail zones, and as such benefit from within-zone cheaper fares. Extending the ticketing dataset to cover the West Yorkshire and the Humber region (as opposed to the West Yorkshire county studies in the current research) will facilitate such analysis. The ATOC have indicated interest in such research and are positively disposed to releasing the relevant ticketing data. This study will require commensurate NRTS data of similar geographic coverage and the DfT have indicated willingness to support such effort. The methodologies developed in this research are suitable and directly applicable to such analysis. The rail sector are keen to incorporate ‘boundary rail-heading’ models into the fare pricing models, and an understanding is needed of the mobility mechanism and categories of passengers who travel further to rail boundaries, in order to benefit from within-zone lower fares. This strand of research fits in as an extension to current rail-heading which is restricted to flows within the West Yorkshire County.

[101] Recall that rail-heading is the name given to the phenomena whereby passengers travel further to access the rail service, when there are commensurately closer access points to the railways.

Another data-driven area of research concerns incorporating PTE[102] tickets into the mobility analysis framework developed during the current research. PTE ticket sales account for about 25% of ticket sales in the West Yorkshire County, and represent a substantial influence on the mobility dynamics recorded in the county. The ATOC are keen that the PTE are incorporated by infilling the RSP ticket database with these PTE tickets. This will give more comprehensive estimates of passenger demand, and by implementation within the urban mobility and movement patterns framework developed in the research, more comprehensive disaggregate demand estimates would be ascertained. There is the further opportunity to assign flows associated with PTE tickets to the network using multi-agent utility maximizing traffic assignment strategies (Axhausen and Gärling, 1992, Horni et al., 2016). These results can be reconciled with traditional logical rules based assignment strategies currently used in the rail industry (SDG, 2012, SDG, 2016). Current on-going discussions about potential availability of PTE tickets have recorded progress at WYCA-ITA.

9.4.2 Spatial microsimulation methods

A shortfall of deterministic spatial microsimulation strategies is that they traditionally will fit a seed data to the marginal of an aggregate array target data. Fitting to the marginal distribution compromises the accuracy of the methodology. A preference would be to fit the seed data to the full joint distribution of the target array. In the current research, the multi-level simultaneous constraint created from the seed data inherently facilitates fitting the seed to the full conditional distribution of the target. This is the case since the target was the aggregate Census interaction data, which is released as 3D (origin-destination versus a socio-demographic attribute) tables. Converting the seed into a 3-level simultaneous hierarchy in effect converts it into the same dimension as the 3D full joint distribution of the target. This was the basis for achieving accurately distributed simulation populations using the methodology developed in the research. However, a limitation of this method would be in scenarios where the target array are of many multiple dimensions, and it then becomes cumbersome to similarly convert the seed to many multi-level simultaneous constraints, with the attendant effective multiplicative reduction in sample size.

[102] Recall PTE tickets are Passenger Transport Executives tickets. PTE’s like the WYCA-ITA manage multi-modal (both bus and train) ticket sales in regional conurbations outside Greater London.

An alternative strategy would be to develop Bayesian methods for reconciling seed data when the target arrays are of very high dimensions. The Bayesian literature (Lunn et al., 2012) includes development on advanced synthesis and meta-analysis strategies which are advantageous as the dimensions of the aggregate target becomes high. Some preliminary effort at developing such synthesis methods have been reported in the literature (Farooq et al., 2013). However, such efforts have not been validated for application to consumer datasets which have been known to be inherently skewed relative to the wider population. The development of population synthesis methods for application to consumer data and social media data is as such a potential area for further research. This would resolve questions regarding whether stochastic simulation based population synthesis methods (typically developed in transportation studies) are comprehensively superior to deterministic spatial micro-simulation methods (typically developed in population geography studies).

9.4.3 GIS-GTFS applications

Front-end applications are a logical step forward to enable non-specialists engage with the technology developed. The development of such front-end applications could form part of future work for the future.

The GIS-GTFS model of mobility developed as part of the research enables the animation of individual rail passengers and individual trains on the rail network. The GIS-GTFS model creates information on the full space-time volume of representative synthetic passengers on the railway network. This includes information on volumes within trains, on platforms, and within train terminals. This model was developed on the commercial ArcGIS platform which is traditional for GIS educational uses. The model developed can be used as an educational tool to illustrate the extents of performance achievable from the educational ArcGIS software platform. In addition, the GIS-GTFS model has potential if reproduced using the Java IDE (to speed up data processing and make the output open-source). Such Java open-source application can be served to passenger mobile phones to visualize predicted micro-level flow volumes of passengers throughout the day, on a chosen day of the week, or when there has been a timetable change during fare rounds. Passengers can as such plan journeys accordingly. This concept will be equivalent to visualising the ‘weather’ of passenger demand in trains, platforms and in train terminals and such applications are targeted at improving customer satisfaction.

9.4.4 Bayesian framework improvements

The Bayesian modelling framework holds promise as a tool for the analysis of complex phenomena like urban mobility. The accuracy of the methodology depends on the efficiency of the chains designed to optimize the solutions. In many instances also, the choice of priors and initial conditions required for objective solutions to be achieved is to an extent still an art. However, there are opportunities for broadening the application of the Bayesian methodology to fully exploit the potentials of consumer data and so called big data.

9.4.4.1 TS-CS nature of big data

The distinctive time-series cross-sectional (TS-CS) nature of the railways ticketing data is not exploited during the research, rather only the cross-section (CS) nature of the data is analysed. As such there exists an opportunity to exploit the TS-CS nature of the ticketing data, which is a particular advantage of big data. The Bayesian modelling framework are suited to the analysis of hierarchical phenomena associated with longitudinal data, which better reveal development trends and the mechanism of mobility change on the railways.

The LENNON ticketing data is captured in periods lasting 4 weeks each. This provides a longitudinal (non-panel) dataset suitable for investigating the time trend of mobility on the UK railways. Basis functions have been employed in literature (Broomhead and Lowe, 1988), but with ticketing data these methods can be used within a Bayesian framework to model time variations in mobility across cross-sectional datasets to reflect seasonal changes in passenger demand due to the range of time-dependent endogenous and exogenous influences. Bayesian convoluted moving average models as proposed in the literature (Ickstadt and Wolpert, 1997), can be used to explore the TS-CS ticketing data. These methods have not been widely explored.

9.4.4.2 Sensitivity of Bayesian models

In traditional Bayesian models the predictive accuracy is achieved by a combination of a sensitivity analysis on the priors and the use of various information criteria to achieve a balance between a parsimonious simple model and a variable laden complete model (McElreath, 2015). In the case of missing data, there are no strategies to before-hand establish the missing data mechanism (i.e. whether it is MAR or not) (Little, 1988), so the modelling strategy adopted is to include as many relevant variables such as to potentially explain the difference between the observed and missing data

(Collins et al., 2001). The many data points in a big data scenario however, could accentuate the noise in the data, introducing higher variability in parameter estimates. There exists the opportunity to develop a robust framework for conducting sensitivity analysis for Bayesian models, as at the moment the Bayesian exercise is as well an art as a science. Bayesian models can take substantial running times, and this is a potential area for a particular breakthrough by improving methodologies that aid in anticipation convergence times of Bayesian models.

9.4.5 Additional agenda

An aggregation of the flows predicted in this research are an order of magnitude lower than those reported from the MOIRA (ORR, 2014c). A future availability of disaggregated MOIRA results will enable a more robust comparison with results from the Bayesian models. An in-house rail industry project can facilitate the comparison of the disaggregate individual-level results from this research with the flows derived from MOIRA. This would provide a fora to validate MOIRA and make suggestions for the improvement of ticketing data captured within the ATOC RSP.

A challenge with the m-IPF is clustering of simulated population when the seed is a low sample ratio. This presents an opportunity to integrate further reference data to improve the variability in the seed. The large consumer data used in this research to a large extent alleviated this problem, but in standard applications low sample ratios could severely hamper the quality of the simulation results. Data collected through GPS tracking of samples of passengers can complement the seed data and alleviate the issue of clustering in the micro-simulated populations. The GPS data would amount to a revealed version of the stated NRTS, and can further serve as validation data to be integrated directly into the spatial microsimulation stage of the framework developed. As the NRTS is a representative sample, the GPS sample procurement would not necessarily need to be facilitated by installing devices on a representative population to be effectively harnessed. Such passenger GPS tracks were not available to the project.

The developed framework potentially serves the growing data assimilation interests for Agent Based Models (ABM’s) of passenger mobility. The rich micro-level passenger attributes and activity can guide the behaviour assigned to agents, and the global attributes identified from regressing the passenger attributes and volumes, would guide in identifying parameters to monitor in the evolution of phenomena in ABM’s.

9.5 Limitations of the work

There have been some limitations in the results derived from the analysis on the consumer ticketing data from the railways. These are discussed in more detail in the following sections. First, the full breath of mobility phenomena in West Yorkshire could not be explored since the flows were limited to those originating and terminating in West Yorkshire. As such a range of interesting mobility phenomena that occur across the boundary could not be explored. Another limitation has been that the hierarchical spatial microsimulation introduced in this thesis to incorporate interaction in the mobility simulation process has some drawbacks. An increase in hierarchy results in an increase in dimensionality of the mobility. This in turn effectively reduces the sample ratio (relative number of sample points), and the attendant consequences on accuracy of results. Additional limitations in the results arise from the NRTS low sample ratios, albeit being the most comprehensive survey in its class. The low sample ratios result in the clustering of the simulated population for many zones. These shortfalls impose limitations on the inferences from the research as discussed below.

9.5.1 Flow limited to within West Yorkshire

The case study presented in this research, is based on flows emanating from and ending in West Yorkshire. As such interregional flows are excluded, implying that about 60% of actual flows have been analysed. Apart from not yielding a comprehensive picture of mobility in the county, this may affect some interesting boundary phenomena like rail-heading, whereby passengers travel further afield to access the rail service, in order to restrict travel to one zone and benefit from cheaper within-zone fares. The rail-heading model developed in this research as such did not account for boundary rail-heading.

A more comprehensive study would include both inter and across region flows, to fully exploit the increased accuracies achievable from the representative micro-level synthetic passenger information. The wider study would achieve a more objective identification of passenger volumes and mobility phenomena. In addition, tickets sold by regional transport executives (PTEs) have not been included, and these amount to ~25% of all flows (SDG, 2014). An exclusion of the PTE ticket sales as such does not reflect a full picture of urban mobility on the rail network within the West Yorkshire study area. This limitation needs to be borne in mind in the interpretation of the results deduced from analysis presented in this thesis.

9.5.2 Clumpy simulated populations

The disaggregation of mobility interaction (into origin-destination pairs) effectively increases the dimensionality of the sample seed. This requires a commensurate increase in volume of sample seed to ensure populated contingency matrices, in-turn ensuring the integrity of any simulated population. The analysis presented in this research is a two-stage spatial microsimulation whereby the NRTS data are used as seed in one of the stages. The relatively small sample ratio of the NRTS results in clusters of individuals with exactly the same attributes created within the simulated population. This represents the so called clumpy data in the simulated population. Although clumpy data are a widespread phenomenon in spatial microsimulation, there are no indices for measuring its extent. The LENNON ticketing data made available to the project were for confidentiality, anonymized by aggregating ticket sales into monthly periods. This aggregation further increases the extent of clumpy data. The NRTS used for simulating a representative population is to date the most comprehensive reference survey of railways passengers in the UK. With the growing number and variety of railway passengers, the NRTS has been reported (ORR, 2014a) to suffer shortfalls due to the lowering sample ratios. These limitations ought to be borne in mind in a bid to manage the expectations of predictive accuracies of the simulated population.

9.5.3 Modifiable accuracy in different zones

The NRTS sample of about 70,000 for West Yorkshire represents a sample ration of 3.5% of the about 2.2 m population. The difference in zonal sizes implies an associated MAUP sample ratio variation across the county. The small sample ratios in some zones imply that there are regions of high replication of the NRTS seed. In areas like Huddersfield (population ~136,100), the simulated (m-IPF) population of ~102,600, an about 75% synthesized population, reflects that the sample seed lacks about 25% of the variability in the Census. In effect, 25% of the unique types of people in the Census for Huddersfield were not captured in the NRTS survey. In the area of Leeds (population ~900,800) the simulated population was ~900,400, an about ~99% of the full variability in the Leeds population. Despite NRTS capturing full Census variability for the area of Leeds, the simulated population results showed that of the 900,400 simulated Leeds population, there are 8,570 (i.e. ~1%) unique individuals, with replication of individuals ranging from 1 to 2856. These issues and limitations have to be borne in mind when using simulated populations from the different geographic zones.

9.5.4 Practical limitations of methodologies

The method of spatial microsimulation introduced in this research extends the conventional IPF optimisation routines established in geography and transport. This is achieved by increasing the cross tabulation of the seed by attaching the origin and destination to each variable. By so doing, the 3D Census target information is now reconciled with the cross tabulated seed. As such the full conditional distribution of the target is fit to the seed, instead of just the marginal in traditional IPF. This modification ensures that the simulated population created not only has an appropriately low TAE, but more importantly has a population distribution of the target population. This also facilitates population synthesis when the seed is skewed relative to the target. However, a limitation of this method would be in scenarios where the target array are of many multiple dimensions, and it then becomes cumbersome to similarly convert the seed to many simultaneous multi-levels. This would yield an effective multiplicative reduction in sample size, thus increasing the lumpiness of the simulated population.

As discussed earlier, Bayesian strategies enable the construction of parsimonious models of complex phenomena. The promise of Bayesian methods hinges on an ability to specify sampling chains that efficiently optimize a solution. In practical situations however, prior and their parameters, as well as initial conditions have to be appropriately specified. For complex problems the analyst does not have an educated intuition about the solution, thereby making it a challenge to specify priors, and initial conditions. These stages of the modelling process can take considerable effort to fulfil, and therein lies the bottleneck in developing Bayesian model solution. The idiosyncrasies in potential strategies for overcoming these bottlenecks, makes Bayesian modelling in its current form an art as much as it is a science. However, the potential of modelling complex phenomena while exploiting the veracity of novel consumer data, achievable within the Bayesian framework makes the challenge worthwhile and holds promise.

Another challenge in Bayesian models is that they can take substantial run times, with the attendant non-guaranteed convergence. A plethora of issues can cause these ranging from specification of the models, the MCMC chains, the prior specifications, and initial conditions. These are potential areas for particular breakthroughs by improving these stages of the methodology to create tools that enable the analyst to anticipate the convergence characteristics and run-times of Bayesian models.

9.6 Future application of developed tools

This research simulated a representative and attribute-rich micro-mobility population of railway passengers embedded in the wider population. The simulated population can serve as input to multi-agent transport modelling tools, and as input to strategic, tactical and operational rail industry planning models. The methodology developed is applicable to a wide variety of typical consumer data which inherently form a skewed subset of the wider population. In the research, the datasets combined were the 2011 Census interaction data (ONS, 2013), the National Rail Travel Survey (NRTS) (DfT, 2013a) and 'big data' consisting of every railway ticket sold in the West Yorkshire study area. However, the methodology developed in this research is scalable and generalizable to other rail transport networks in areas within the UK, or in other countries where large consumer revealed datasets and similar background stated choice surveys are available.

The GIS-GTFS model is potentially useful to network operators for management, maintenance and interventions assessment on the UK railways. The representative micro-level population can aid in creating segmentations on railway customers satisfied by particular service implementations. Such segregation can aid in answering pertinent policy questions related to how concessions might affect different socio-demographic groups within populations. The GIS-GTFS network model developed can be used as a tool for visualising or animating the mobility of simulated passengers and transit trains on the railway network. Such an application can be served to potential railway customers to facilitate journey planning.

The Bayesian modelling framework provides a tool for the imputation of passenger behavioural attributes, and is applicable in establishing comprehensive volumes of passenger demand, and drivers of mobility phenomena on the railways.

List of References

ACUNA, E. & RODRIGUEZ, C. 2004. The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications.* Springer.

ADAMOWICZ, W., LOUVIERE, J. & WILLIAMS, M. 1994. Combining Revealed and Stated Preference Methods for Valuing Environmental Amenities. *Journal of Environmental Economics and Management,* 26**,** 271-292.

ADOLPH, S., COCKBURN, A. & BRAMBLE, P. 2002. *Patterns for effective use cases*, Addison-Wesley Longman Publishing Co., Inc.

AKAIKE, H. 1987. Factor analysis and AIC. *Psychometrika,* 52**,** 317-332.

ALBERT, J. 2009. *Bayesian computation with R*, Springer Science & Business Media.

ALLISON, P. D. 2001. *Missing data*, Sage publications.

ALLISON, P. D. Handling missing data by maximum likelihood. SAS global forum, 2012. Statistical Horizons, Havenford, PA.

AMOS, P., BULLOCK, D. & SONDHI, J. 2010. High-speed rail: The fast track to economic development. *The World Bank***,** 1-28.

ANDERSSON, J. 2017. Using Geographically Weighted Regression (GWR) to explore spatial variations in the relationship between public transport accessibility and car use: a case study in Lund and Malmö, Sweden. *Student thesis series INES*.

ANDERSTIG, C. & MATTSSON, L.-G. 1998. Modelling land-use and transport interaction: policy analyses using the IMREL model. *Network Infrastructure and the Urban Environment.* Springer.

ANGRIST, J. D., IMBENS, G. W. & RUBIN, D. B. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association,* 91**,** 444-455.

ANTHONY, R. N. 1965. *Planning and Control Systems: A Framework for Analysis [by]*, Division of Research, Graduate School of Business Administration, Harvard University.

ANTRIM, A. & BARBEAU, S. J. 2013. The many uses of GTFS data–opening the door to transit and multimodal applications. *Location-Aware Information Systems Laboratory at the University of South Florida***,** 4.

ARCHIVE. 2011. *Managing and Sharing Data - UK Data Archive* [Online]. Available: www.data-archive.ac.uk/media/2894/managingsharing.pdf [Accessed 9th March 2015 2015].

ARON, A. & ARON, E. N. 1994. *Statistics for psychology*, Prentice-Hall, Inc.

ASMUSSEN, S. & GLYNN, P. W. 2007. *Stochastic simulation: algorithms and analysis*, Springer Science & Business Media.

ASSAD, A. A. 1980. Models for rail transportation. *Transportation Research Part A: General,* 14**,** 205-220.

ATOC. 2003a. *New Settlement System Goes Live Across The Uk Rail Network* [Online]. http://www.atoc.org/: ATOC. [Accessed December 28 2015].

ATOC 2003b. New Settlement System Goes Live Across The Uk Rail Network. 28 August 2003 ed.: Association of Train Operating Companies.

ATOC, D., RSP 2017. GB Rail GTFS feed.

AXHAUSEN, K. W. & GÄRLING, T. 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews,* 12**,** 323-341.

AZUR, M. J., STUART, E. A., FRANGAKIS, C. & LEAF, P. J. 2011. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research,* 20**,** 40-49.

BACHARACH, M. 1970. *Biproportional matrices and input-output change*, CUP Archive.

BAJRAM SPASESKI, N. & LER, D. 2013. *Application of Discrete-Time Markov Models*.

BALLAS, D., CLARKE, G., DORLING, D., EYRE, H., THOMAS, B. & ROSSITER, D. 2005. SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place,* 11**,** 13-34.

BALLAS, D. & CLARKE, G. P. 2001. Modelling the local impacts of national social policies: a spatial microsimulation approach. *Environment and Planning C: Government and Policy,* 19**,** 587-606.

BALLAS, D., KINGSTON, R., STILLWELL, J. & JIN, J. 2007. Building a spatial microsimulation-based planning support system for local policy making. *Environment and Planning A,* 39**,** 2482-2499.

BARTHELEMY, J., SUESSE, T., NAMAZI-RAD, M. & BARTHELEMY, M. J. 2016. Package ‘mipfp’.

BARTHELEMY, J. & SUESSE, T. F. 2016. Package mipfp: Multidimensional Iterative Proportional Fitting and Alternative Models. R package version 3.1.

BARTHÉLEMY, M. 2011. Spatial networks. *Physics Reports,* 499**,** 1-101.

BARTLETT, J. W., CARPENTER, J. R., TILLING, K. & VANSTEELANDT, S. 2014. Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics,* 15**,** 719-730.

BATISTA, G. E. & MONARD, M. C. 2002. A Study of K-Nearest Neighbour as an Imputation Method. *HIS,* 87**,** 48.

BATTY, M., AXHAUSEN, K. W., GIANNOTTI, F., POZDNOUKHOV, A., BAZZANI, A., WACHOWICZ, M., OUZOUNIS, G. & PORTUGALI, Y. 2012. Smart cities of the future. *The European Physical Journal Special Topics,* 214**,** 481-518.

BATTY, M. & MACKIE, S. 1972. The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning A,* 4**,** 205-233.

BAYARRI, M. J. & BERGER, J. O. 2004. The interplay of Bayesian and frequentist analysis. *Statistical Science***,** 58-80.

BAYES, T., PRICE, R. & CANTON, J. 1763. An essay towards solving a problem in the doctrine of chances.

BECK, N. 2008. Time-series-cross-section methods. *Oxford handbook of political methodology***,** 475-93.

BECK, N. & KATZ, J. N. 1995. What to do (and not to do) with time-series cross-section data. *American political science review,* 89**,** 634-647.

BEN-AKIVA, M., RAMMING, M. S. & BEKHOR, S. 2004. Route choice models. *Human Behaviour and Traffic Networks***,** 23-46.

BEN-AKIVA, M. E. & LERMAN, S. R. 1985a. *Discrete choice analysis: theory and application to travel demand*, MIT press.

BEN-AKIVA, M. E. & LERMAN, S. R. 1985b. *Discrete choice analysis: theory and application to travel demand,* Cambridge, Mass; London, MIT Press.

BERTSEKAS, D. P. 2014. *Constrained optimization and Lagrange multiplier methods*, Academic press.

BERTSIMAS, D. 1988. *Probabilistic combinatorial optimization problems.* Massachusetts Institute of Technology.

BESAG, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 192-236.

BESAG, J., YORK, J. & MOLLIÉ, A. 1991. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics,* 43**,** 1-20.

BEŠINOVIĆ, N., QUAGLIETTA, E. & GOVERDE, R. M. 2013. A simulation-based optimization approach for the calibration of dynamic train speed profiles. *Journal of Rail Transport Planning & Management,* 3**,** 126-136.

BEST, N. G., ICKSTADT, K. & WOLPERT, R. L. 2000. Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American statistical association,* 95**,** 1076-1088.

BHAT, C. R. & KOPPELMAN, F. S. 1999. Activity-based modeling of travel demand. *Handbook of transportation Science.* Springer.

BIFULCO, G. N., CARTENÌ, A. & PAPOLA, A. 2010. An activity-based approach for complex travel behaviour modelling. *European transport research review,* 2**,** 209-221.

BIRKIN, M. & CLARKE, G. 1995. Using microsimulation methods to synthesize census data. *Census users' handbook***,** 363-387.

BIRKIN, M., CLARKE, G. & CLARKE, M. 2010. Refining and Operationalizing Entropy-Maximizing Models for Business Applications. 商业应用模式下熵最大化模型的应用与改进. *Geographical Analysis,* 42**,** 422-445.

BIRKIN, M. & CLARKE, M. 1988. SYNTHESIS—a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and planning A,* 20**,** 1645-1671.

BIRKIN, M. & CLARKE, M. 1989. The generation of individual and household incomes at the small area level using synthesis. *Regional Studies,* 23**,** 535-548.

BIRKIN, M. & CLARKE, M. 2011. Spatial microsimulation models: A review and a glimpse into the future. *Population Dynamics and Projection Methods.* Springer.

BIRKIN, M., TURNER, A. & WU, B. A synthetic demographic model of the UK population: Methods, progress and problems. Regional Science Association International British and Irish Section, 36th Annual Conference, 2006.

BISHOP, Y. M., FIENBERG, S. E. & HOLLAND, P. W. 2007. *Discrete multivariate analysis: theory and practice*, Springer Science & Business Media.

BIVAND, R. S., PEBESMA, E. J., GÓMEZ-RUBIO, V. & PEBESMA, E. J. 2008. *Applied spatial data analysis with R*, Springer.

BOLLEN, J., MAO, H. & ZENG, X. 2011. Twitter mood predicts the stock market. *Journal of computational science,* 2**,** 1-8.

BOWMAN, J. L. & BEN-AKIVA, M. E. 2001. Activity-based disaggregate travel demand model system with activity schedules. *Transportation research part a: policy and practice,* 35**,** 1-28.

BOX, G. E. & TIAO, G. C. 2011. *Bayesian inference in statistical analysis*, John Wiley & Sons.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. & OLSHEN, R. A. 1984. *Classification and regression trees*, CRC press.

BRODERSEN, K. H., GALLUSSER, F., KOEHLER, J., REMY, N. & SCOTT, S. L. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics,* 9**,** 247-274.

BROOKS, S., GELMAN, A., JONES, G. & MENG, X.-L. 2011. *Handbook of Markov Chain Monte Carlo*, CRC press.

BROOMHEAD, D. S. & LOWE, D. 1988. Radial basis functions, multi-variable functional interpolation and adaptive networks. Royal Signals and Radar Establishment Malvern (United Kingdom).

BROWNSTONE, D. 2001. Discrete choice modeling for transportation. *University of California Transportation Center*.

BRUNSDON, C., FOTHERINGHAM, A. S. & CHARLTON, M. E. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis,* 28**,** 281-298.

BUSCHER, V. D., L.; WEBB, M.; AOUN, C. 2014. Urban Mobility in the Smart City Age. *In:* EBI, K. (ed.) *Smart Cities Cornerstone Series.* http://smartcitiescouncil.com/resources/urban-mobility-smart-city-age: Arup, The Climate Group, Schneider Electric.

BUTLER, D. 2013. When Google got flu wrong. *Nature,* 494**,** 155.

BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software,* 45.

BYRNE, T. 2000. *Local government in Britain: everyone's guide to how it all works*, Penguin.

CADWALLADR, C. & GRAHAM-HARRISON, E. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian,* 17.

CARDNO, A. J. & MULGAN, N. J. 2003. Travel route planner system and method. Google Patents.

CARLIN, B. P., GELFAND, A. E. & SMITH, A. F. 1992. Hierarchical Bayesian analysis of changepoint problems. *Applied statistics***,** 389-405.

CHAPIN, F. S. 1974. *Human activity patterns in the city: things people do in time and in space,* New York; London, Wiley.

CHIB, S. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association,* 90**,** 1313-1321.

CHIB, S. & GREENBERG, E. 1995. Understanding the metropolis-hastings algorithm. *The american statistician,* 49**,** 327-335.

CLARKE, F. H. 1990. *Optimization and nonsmooth analysis*, SIAM.

CLAUSET, A. 2011. A brief primer on probability distributions. *Santa Fe Institute. http://tuvalu. santafe. edu/~ aaronc/courses/7000/csci7000-001_2011_L0. pdf*.

CLAYTON, D. & KALDOR, J. 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics***,** 671-681.

CLEVELAND, W. S. & DEVLIN, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association,* 83**,** 596-610.

CODD, E. F. 1970. A relational model of data for large shared data banks. *Communications of the ACM,* 13**,** 377-387.

COLLINS, L. M., SCHAFER, J. L. & KAM, C.-M. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods,* 6**,** 330.

CONGDON, P. 2014. *Applied bayesian modelling*, John Wiley & Sons.

CONSULTANTS, I. 1999. Emme/2 User" s Manual: Release 9.2. *Montréal, Canada*.

COX, D. R. & SNELL, E. J. 1968. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 248-275.

CRESSIE, N. 1990. The origins of kriging. *Mathematical geology,* 22**,** 239-252.

DALGAARD, P. 2008. Rates and Poisson regression. *Introductory Statistics with R.* Springer.

DANGERMOND, J. 1992. What is a Geographic Information System (GIS)? *Geographic Information Systems (GIS) and Mapping—Practices and Standards.* ASTM International.

DAVIS, T. E. & PRINCIPE, J. C. 1993. A Markov chain framework for the simple genetic algorithm. *Evolutionary computation,* 1**,** 269-288.

DE DIOS ORTUZAR, J. & WILLUMSEN, L. G. 1994. *Modelling transport*, Wiley New Jersey.

DE FINETTI, B. 1972. *Probability, induction and statistics: the art of guessing,* London, J. Wiley.

DE JONG, G., DALY, A., PIETERS, M., MILLER, S., PLASMEIJER, R. & HOFMAN, F. 2007. Uncertainty in traffic forecasts: literature review and new results for The Netherlands. *Transportation,* 34**,** 375-395.

DE JONG, G., GUNN, H. & WALKER, W. 2004. National and international freight transport models: an overview and ideas for future development. *Transport Reviews,* 24**,** 103-124.

DEMING, W. E. & STEPHAN, F. F. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics,* 11**,** 427-444.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** 1-38.

DEMPSTER, M. A. 2012. *Deterministic and Stochastic Scheduling: Proceedings of the NATO Advanced Study and Research Institute on Theoretical Approaches to Scheduling Problems held in Durham, England, July 6–17, 1981*, Springer Science & Business Media.

DENNETT, A. 2012. Estimating flows between geographical locations:'get me started in'spatial interaction modelling. *UCL working paper series,* 181**,** 1-24.

DFT 2013a. National Rail Travel Survey – Overview report. *In:* TRANSPORT, D. F. (ed.) *Rail statistics.* www.gov.uk: Department for Transport.

DFT. 2013b. *National Travel Survey: England 2013* [Online]. Department for Transport. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/342160/nts2013-01.pdf [Accessed 12 December 2017].

DIGGLE, P. J., TAWN, J. & MOYEED, R. 1998. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 47**,** 299-350.

DIMITRIOU, L., TSEKERIS, T. & STATHOPOULOS, A. 2006. Genetic-algorithm-based micro-simulation approach for estimating turning proportions at signalized intersections. *IFAC Proceedings Volumes,* 39**,** 159-164.

DOCHERTY, M. & SMITH, R. 1999. The case for structuring the discussion of scientific papers: Much the same as that for structuring abstracts. *BMJ: British Medical Journal,* 318**,** 1224.

DOMINICI, F., MCDERMOTT, A., ZEGER, S. L. & SAMET, J. M. 2002. On the use of generalized additive models in time-series studies of air pollution and health. *American journal of epidemiology,* 156**,** 193-203.

DONDERS, A. R. T., VAN DER HEIJDEN, G. J., STIJNEN, T. & MOONS, K. G. 2006. Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology,* 59**,** 1087-1091.

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**,** 163-180.

DONG, Y. & PENG, C.-Y. J. 2013. Principled missing data methods for researchers. *SpringerPlus,* 2**,** 1-17.

DOWLE, M., SRINIVASAN, A., GORECKI, J., SHORT, T., LIANOGLOU, S., ANTONYAN, E. & DOWLE, M. M. 2017. Package ‘data. table’. *R package version,* 1.

DUAN, Y., BREHM, W., STROBL, H., TITTLBACH, S., HUANG, Z. & SI, G. 2013. Steps to and correlates of health-enhancing physical activity in adulthood: an intercultural study between German and Chinese individuals. *Journal of Exercise Science & Fitness,* 11**,** 63-77.

DUGUAY, G., JUNG, W. & MCFADDEN, D. 1976. *SYNSAM: A methodology for synthesizing household transportation survey data*, Urban Travel Demand Forecasting Project, Institute of Transportation Studies.

DUNCAN, O. D. 1961. A socioeconomic index for all occupations. *Class: Critical Concepts,* 1**,** 388-426.

DURRANT, G. B. 2005. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *National Center for Research Methods Working Paper,* 2.

EDINBURGHDATASHARE 2016. Research Data Service, University of Edinburgh. Edinburgh Data Share.

EDWARDS, K. L., CLARKE, G. P., THOMAS, J. & FORMAN, D. 2011. Internal and external validation of spatial microsimulation models: small area estimates of adult obesity. *Applied Spatial Analysis and Policy,* 4**,** 281-300.

EFRON, B. 1994. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association,* 89**,** 463-475.

EFRON, B. & TIBSHIRANI, R. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science***,** 54-75.

EGENHOFER, M. & KUHN, W. 1999. Interacting with GIS. *Geographical Information Systems: Principles, techniques, applications, and management,* 1**,** 401-412.

ELLISON, N. B., STEINFIELD, C. & LAMPE, C. 2007. The benefits of Facebook “friends:” Social capital and college students’ use of online social network sites. *Journal of computer-mediated communication,* 12**,** 1143-1168.

ENDERS, C. K. & BANDALOS, D. L. 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling,* 8**,** 430-457.

EPSTEIN, J. M. & AXTELL, R. 1996. *Growing artificial societies: social science from the bottom up*, Brookings Institution Press.

ESCOBAR, M. D. & WEST, M. 1995. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association,* 90**,** 577-588.

ESRI 2016. ArcGIS, 10.3. *Redlands, California: ESRI*.

ETTEMA, D. & TIMMERMANS, H. 1997. *Activity-based approaches to travel analysis:[selected papers presented at the workshop, May 1995]*.

EVERETT III, H. 1963. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research,* 11**,** 399-417.

EVERITT, B. 2012. *Introduction to optimization methods and their application in statistics*, Springer Science & Business Media.

EWING, G. 1974. Gravity and linear regression models of spatial interaction: a cautionary note. *Economic Geography,* 50**,** 83-88.

FAROOQ, B. & BIERLAIRE, M. Simulation-based population synthesis using Gibbs sampling. Conference on synthetic population, 2017.

FAROOQ, B., BIERLAIRE, M., HURTUBIA, R. & FLÖTTERÖD, G. 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological,* 58**,** 243-263.

FERMAT, P. D. 1891. Oeuvres de Fermat, ed. *Charles Henry and Paul Tannery,* 5**,** 1891-1922.

FIENBERG, S. E. 1970. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics***,** 907-917.

FISHER, R. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A,* 222**,** 309-368.

FLETCHER, R. 2013. *Practical methods of optimization*, John Wiley & Sons.

FOTHERINGHAM, A. S. 1983. A new set of spatial-interaction models: the theory of competing destinations. *Environment & Planning A.*

FOTHERINGHAM, A. S., BRUNSDON, C. & CHARLTON, M. 2003. *Geographically weighted regression: the analysis of spatially varying relationships*, John Wiley & Sons.

FOTHERINGHAM, A. S., CHARLTON, M. E. & BRUNSDON, C. 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A,* 30**,** 1905-1927.

FOWKES, A., NASH, C. & WHITEING, A. 1985. Understanding trends in inter-city rail traffic in Great Britain. *Transportation Planning and Technology,* 10**,** 65-80.

FRATAR, T. J. 1954. Vehicular trip distribution by successive approximations. *Traffic Quarterly,* 8.

FROME, E. L. 1983. The analysis of rates using Poisson regression models. *Biometrics***,** 665-674.

FUCHS, V. R. 2004. Reflections on the socio-economic correlates of health. *Journal of health economics,* 23**,** 653-661.

FURNESS, K. 1965. Time function iteration. *Traffic Engineering and Control,* 7**,** 458-460.

GALILI, T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics,* 31**,** 3718-3720.

GAO, W., BALMER, M. & MILLER, E. 2010. Comparison of MATSim and EMME/2 on greater Toronto and Hamilton area network, Canada. *Transportation Research Record: Journal of the Transportation Research Board***,** 118-128.

GARCIA, R. I., IBRAHIM, J. G. & ZHU, H. 2010. Variable selection for regression models with missing data. *Statistica Sinica,* 20**,** 149.

GARDNER, W., MULVEY, E. P. & SHAW, E. C. 1995. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological bulletin,* 118**,** 392.

GAY, P. 1966. *Age of enlightenment*.

GELFAND, A. E., HILLS, S. E., RACINE-POON, A. & SMITH, A. F. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association,* 85**,** 972-985.

GELMAN, A. 2007. Statistical modeling, causal inference and social science. *URL http://www. stat. columbia. edu/gelman/blog*.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. 2014a. *Bayesian data analysis*, CRC press Boca Raton, FL.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. 2014b. *Bayesian data analysis*, Chapman & Hall/CRC Boca Raton, FL, USA.

GELMAN, A. & HILL, J. 2006. *Data analysis using regression and multilevel/hierarchical models*, Cambridge university press.

GEMAN, S. & GEMAN, D. 1987. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Readings in Computer Vision.* Elsevier.

GEORGE, G., HAAS, M. R. & PENTLAND, A. 2014. Big data and management. *Academy of Management Journal,* 57**,** 321-326.

GEYER, C. J. 1992. Practical markov chain monte carlo. *Statistical science***,** 473-483.

GIANNOTTI, F., NANNI, M., PEDRESCHI, D., PINELLI, F., RENSO, C., RINZIVILLO, S. & TRASARTI, R. 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases,* 20**,** 695-719.

GILKS, W. R., RICHARDSON, S. & SPIEGELHALTER, D. 1995. *Markov chain Monte Carlo in practice*, CRC press.

GLEAVE, S. D. 2011. The value of station investment: research on regenerative impacts. *London: Network Rail*.

GONZALEZ, M. C., HIDALGO, C. A. & BARABASI, A.-L. 2008. Understanding individual human mobility patterns. *Nature,* 453**,** 779-782.

GOODCHILD, M. F. 1992. Geographical information science. *International journal of geographical information systems,* 6**,** 31-45.

GOOGLE. 2016. *Google Transit API* [Online]. https://developers.google.com: Google, Creative Commons Attribution 3.0 License. Available: https://developers.google.com/transit/gtfs/reference/ [Accessed July 26 2017].

GOOGLE. 2017. *Google Maps Distance Matrix API* | *Google Developers* [Online]. Available: https://developers.google.com/maps/documentation/distance-matrix/ [Accessed 30 December 2017].

GOUDIE, R. J. & MUKHERJEE, S. 2016. A Gibbs sampler for learning DAGs. *The Journal of Machine Learning Research,* 17**,** 1032-1070.

GOWER, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics***,** 857-871.

GRACE-MARTIN, K. 2018. *The Analysis Factor* [Online]. The Analysis Factor. Available: https://www.theanalysisfactor.com/seven-ways-to-make-up-data-common-methods-to-imputing-missing-data/ [Accessed May 10, 2018 2018].

GRAHAM, J. W. 2003. Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling,* 10**,** 80-100.

GRAHAM, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology,* 60**,** 549-576.

GRAHAM, J. W., HOFER, S. M. & MACKINNON, D. P. 1996. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research,* 31**,** 197-218.

GRAPPERON, A. 2016. *Behavioral Approach to Estimation of Smart Card Holders Socio-Demographic Characteristics in a Public Transportation System.* École Polytechnique de Montréal.

GRAPPERON, A., FAROOQ, B. & TREPANIER, M. 2016. Information fusion of smart card data with travel survey.

GREENE, W. H. 1994. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models.

GRIMMER, J. 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics,* 48**,** 80-83.

GSCHLÖßL, S. & CZADO, C. 2008. Modelling count data with overdispersion and spatial effects. *Statistical papers,* 49**,** 531-552.

GUO, J. & BHAT, C. 2007. Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board***,** 92-101.

HADAYEGHI, A., SHALABY, A. S. & PERSAUD, B. N. 2010. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. *Accident Analysis & Prevention,* 42**,** 676-688.

HÄGERSTRAND, T. 1970. What about people in Regional Science? *Papers of the Regional Science Association,* 24**,** 6-21.

HALD, A. 2008. *A history of parametric statistical inference from Bernoulli to Fisher, 1713-1935*, Springer Science & Business Media.

HALL, M. & WILLUMSEN, L. 1980. SATURN-a simulation-assignment model for the evaluation of traffic management schemes. *Traffic Engineering & Control,* 21.

HARLAND, K., HEPPENSTALL, A., SMITH, D. & BIRKIN, M. 2012. Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. *Journal of Artifical Societies and Social Simulation,* 15**,** 1-15.

HARRELL, F. E., LEE, K. L. & MARK, D. B. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine,* 15**,** 361-387.

HASTINGS, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* 57**,** 97-109.

HC-700-II 2006. How fair are the fares? Train fares and ticketing-Committee's sixth report of session 2005-06-Volume II. *In:* COMMITTEE, G. B. P. H. O. C. T. (ed.) *HC 700-II.*

HECKMAN, J. J. 1977. Sample selection bias as a specification error (with an application to the estimation of labor supply functions). National Bureau of Economic Research Cambridge, Mass., USA.

HECKMAN, J. J. 1979. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society***,** 153-161.

HEITJAN, D. F. & BASU, S. 1996. Distinguishing "missing at random" and "missing completely at random". *The American Statistician,* 50**,** 207-213.

HENRETTY, N. 2011/12. Small area income estimates for middle layer super output areas, England and Wales. 16 December 2016 ed.: Office for National Statistics

HENSHER, D. A. & BULLOCK, R. G. 1979. Price elasticity of commuter mode choice: effect of a 20 per cent rail fare reduction. *Transportation Research Part A: General,* 13**,** 193-202.

HENSHER, D. A. & BUTTON, K. J. 2007. *Handbook of transport modelling*, Emerald Group Publishing Limited.

HEPPENSTALL, A. J., CROOKS, A. T., SEE, L. M. & BATTY, M. 2011. *Agent-based models of geographical systems*, Springer Science & Business Media.

HESS, S. & DALY, A. 2014. *Handbook of choice modelling*, Edward Elgar Publishing.

HO, D. E., IMAI, K., KING, G. & STUART, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis,* 15**,** 199-236.

HONAKER, J. & KING, G. 2010. What to do about missing values in time-series cross-section data. *American Journal of Political Science,* 54**,** 561-581.

HONAKER, J., KING, G. & BLACKWELL, M. 2011. Amelia II: A program for missing data. *Journal of Statistical Software,* 45**,** 1-47.

HORNI, A., NAGEL, K. & AXHAUSEN, K. W. 2016. *The multi-agent transport simulation MATSim*, Ubiquity Press London.

HORTON, N. J. & KLEINMAN, K. P. 2007. Much ado about nothing. *The American Statistician,* 61.

HORTON, N. J. & LAIRD, N. M. 1999. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research,* 8**,** 37-50.

HORTON, N. J. & LIPSITZ, S. R. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician,* 55**,** 244-254.

HOWE, D., COSTANZO, M., FEY, P., GOJOBORI, T., HANNICK, L., HIDE, W., HILL, D. P., KANIA, R., SCHAEFFER, M. & ST PIERRE, S. 2008. Big data: The future of biocuration. *Nature,* 455**,** 47-50.

HUMMEL, M., EDELMANN, D. & KOPP-SCHNEIDER, A. 2017. CluMix: Clustering and Visualization of Mixed-Type Data.

HUYNH, N. N., CAO, V., WICKRAMASURIYA DENAGAMAGE, R., BERRYMAN, M., PEREZ, P. & BARTHELEMY, J. 2014. An agent based model for the simulation of road traffic and transport demand in a Sydney metropolitan area.

IBRAHIM, J. G. & MOLENBERGHS, G. 2009. Missing data methods in longitudinal studies: a review. *Test,* 18**,** 1-43.

ICKSTADT, K. & WOLPERT, R. L. 1997. Multiresolution assessment of forest inhomogeneity. *Case Studies in Bayesian Statistics.* Springer.

ICO. 2015. *What are the Environmental Information Regulations?* [Online]. Information Commissioner's Office. Available: www.ico.org.uk/for-organisations/guide-to-the-environmental-information-regulations/what-are-the-eir/ [Accessed March 9th 2015 2015].

INSPIRE. 2015. *Infrastructure for Spatial Information in the European Community* [Online]. INSPIRE. Available: http://inspire.ec.europa.eu/ [Accessed 9th May 2015 2015].

IRELAND, C. T. & KULLBACK, S. 1968. Contingency tables with given marginals. *Biometrika,* 55**,** 179-188.

JACCARD, J. & TURRISI, R. 2003. *Interaction effects in multiple regression*, Sage.

JARA-DÍAZ, S. R. 1998. Time and income in travel demand: towards a microeconomic activity framework. *Theoretical Foundations of Travel Choice Modelling. Elsevier*.

JENSEN, F. V. 1996. *An introduction to Bayesian networks*, UCL press London.

JONES, P. M., DIX, M. C., CLARKE, M. I. & HEGGIE, I. G. 1983. *Understanding travel behaviour*.

JUSTICE. 2008. *Freedom of information guidance Exemptions guidance Section 40 – Personal information* [Online]. Available: www.justice.gov.uk/downloads/information-access-rights/foi/foi-exemption-s40.pdf [Accessed 9th March 2015 2015].

KASS, R. E., CARLIN, B. P., GELMAN, A. & NEAL, R. M. 1998. Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician,* 52**,** 93-100.

KAVROUDAKIS, D. 2015. sms: Microdata for Geographical Analysis in R. *Journal of Statistical Software,* 68**,** 1-23.

KBIOB, D. 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of Chemical, Metallurgical, and Mining Society of South Africa*.

KELLEY, C. T. 1999. *Iterative methods for optimization*, SIAM.

KELSALL, J. & WAKEFIELD, J. 1999. Discussion of ‘Bayesian models for spatially correlated disease and exposure data’, by Best et al. *Bayesian statistics,* 6**,** 151.

KING, G., HONAKER, J., JOSEPH, A. & SCHEVE, K. List-wise deletion is evil: what to do about missing data in political science. Annual Meeting of the American Political Science Association, Boston, 1998.

KING, G., HONAKER, J., JOSEPH, A. & SCHEVE, K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Association, 2001. Cambridge Univ Press, 49-69.

KINGMAN, J. F. 1978. Uses of exchangeability. *The Annals of Probability***,** 183-197.

KITCHIN, R. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society,* 1**,** 2053951714528481.

KNUTTI, R. & SEDLÁCEK, J. 2013. Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change,* 3**,** 369.

KOCH, G., PROBABILITY, I. C. O. E. I., STATISTICS & DE FINETTI, B. 1982. *Exchangeability in probability and statistics: proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th-9th April, 1981, in honour of Professor Bruno de Finetti*, North-Holland Publ.

KRAJZEWICZ, D., HERTKORN, G., RÖSSEL, C. & WAGNER, P. SUMO (Simulation of Urban MObility)-an open-source traffic simulation. Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002), 2002. 183-187.

KRUSCHKE, J. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*, Academic Press.

KUMAR, A. 1980. *Use of incremental form of logit models in demand analysis*.

KUNER, C. 2012. The European Commission's proposed data protection regulation: A copernican revolution in European data protection law.

KURBAN, H., GALLAGHER, R., KURBAN, G. A. & PERSKY, J. 2011. A beginner's guide to creating small-area cross-tabulations. *Cityscape***,** 225-235.

LAGRANGE, J. L. 1867. Oeuvres de Lagrange (14 vols). *Gauthier-Villars, Paris*.

LAKE, L. W. 2016. *Kriging and cokriging* [Online]. petrowiki.org: Society of Petroleum Engineers International. Available: www.petrowiki.org [Accessed 6th January 2016].

LANDERMAN, L. R., LAND, K. C. & PIEPER, C. F. 1997. An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods & Research,* 26**,** 3-33.

LANGE, K. L., LITTLE, R. J. & TAYLOR, J. M. 1989. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association,* 84**,** 881-896.

LAVRAKAS, P. J. 2008. *Encyclopedia of survey research methods*, Sage Publications.

LAWSON, A. B., BROWNE, W. J. & RODEIRO, C. L. V. 2003. *Disease mapping with WinBUGS and MLwiN*, John Wiley & Sons.

LAZER, D., KENNEDY, R., KING, G. & VESPIGNANI, A. 2014. The parable of Google Flu: traps in big data analysis. *Science,* 343**,** 1203-1205.

LEE, D. 2013. CARBayes: an R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software,* 55**,** 1-24.

LEEDS, I. 2006. KONSULT Knowledgebase on Sustainable Urban Land Use and Transport. *Institute for Transport Studies, University of Leeds, Leeds, UK. http://konsult. leeds. ac. uk/public/level0/l0_hom. htmi*.

LEUNG, J. Y. 2004. *Handbook of scheduling: algorithms, models, and performance analysis*, CRC Press.

LITTLE, R. & RUBIN, D. 2002. Statistical Analysis with Missing Data.

LITTLE, R. J. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association,* 83**,** 1198-1202.

LITTLE, R. J. & RUBIN, D. B. 2014. *Statistical analysis with missing data*, John Wiley & Sons.

LOMAX, N. & NORMAN, P. 2015. Estimating Population Attribute Values in a Table:"Get Me Started in" Iterative Proportional Fitting. *The Professional Geographer***,** 1-11.

LOMAX, N. & NORMAN, P. 2016. Estimating Population Attribute Values in a Table:"Get Me Started in" Iterative Proportional Fitting. *The Professional Geographer,* 68**,** 451-461.

LOMAX, N., NORMAN, P., REES, P. & STILLWELL, J. 2013. Subnational migration in the United Kingdom: producing a consistent time series using a combination of available data and estimates. *Journal of Population Research,* 30**,** 265-288.

LONG, J. S. & FREESE, J. 2006. *Regression models for categorical dependent variables using Stata*, Stata press.

LONGLEY, P. 2005. *Geographic information systems and science*, John Wiley & Sons.

LOVELACE, R. & BALLAS, D. 2013. 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems,* 41**,** 1-11.

LOVELACE, R., BIRKIN, M., BALLAS, D. & VAN LEEUWEN, E. 2015. Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. *Journal of Artificial Societies and Social Simulation,* 18**,** 21.

LOVELACE, R. & DUMONT, M. 2016. *Spatial Microsimulation with R*, CRC Press.

LUNN, D., JACKSON, C., BEST, N., THOMAS, A. & SPIEGELHALTER, D. 2012. *The BUGS book: A practical introduction to Bayesian analysis*, CRC press.

LUNN, D. J., THOMAS, A., BEST, N. & SPIEGELHALTER, D. 2000. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing,* 10**,** 325-337.

MALLESON, N. & BIRKIN, M. 2012. Estimating individual behaviour from massive social data for an urban agent-based model. *Modeling Social Phenomena in Spatial Context.* Lit Verlag Berlin.

MANTELERO, A. 2013. The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review,* 29**,** 229-235.

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. & BYERS, A. H. 2011. Big data: The next frontier for innovation, competition, and productivity.

MARCHENKO, Y. Chained equations and more in multiple imputation in Stata 12. 2011 Italian Stata Users Group Meeting, 2011.

MARR, B. 2015. *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*, John Wiley & Sons.

MAZUMDER, R., HASTIE, T. & TIBSHIRANI, R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research,* 11**,** 2287-2322.

MCELREATH, R. 2015. Statistical Rethinking. Texts in Statistical Science. CRC Press.

MCFADDEN, D. 1973. Conditional logit analysis of qualitative choice behavior.

MCFADDEN, D. 1980. Econometric models for probabilistic choice among products. *Journal of Business***,** S13-S29.

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

MCHUGH, B. 2013. Pioneering open data standards: The GTFS story. *Beyond transparency: open data and the future of civic innovation***,** 125-135.

MCMILLEN, D. P. 2004. Geographically weighted regression: the analysis of spatially varying relationships. Oxford University Press.

MCNALLY, M. G. & RINDT, C. R. 2007. The activity-based approach. *Handbook of Transport Modelling: 2nd Edition.* Emerald Group Publishing Limited.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics,* 21**,** 1087-1092.

METROPOLIS, N. & ULAM, S. 1949. The Monte Carlo Method. *Journal of the American Statistical Association,* 44**,** 335-341.

MILLER, E. J., HUNT, J. D., ABRAHAM, J. E. & SALVINI, P. A. 2004. Microsimulating urban systems. *Computers, environment and urban systems,* 28**,** 9-44.

MORANG, M. 2016. *Using GTFS Data in ArcGIS Network Analyst* [Online]. http://transit.melindamorang.com/. Available: http://transit.melindamorang.com/ [Accessed 24 September 2016].

MORRIS, A. E. J. 2013. *History of urban form before the industrial revolution*, Routledge.

MORRIS, P. The breakout method for escaping from local minima. AAAI, 1993. 40-45.

MÜHLENBEIN, H., GORGES-SCHLEUTER, M. & KRÄMER, O. 1988. Evolution algorithms in combinatorial optimization. *Parallel Computing,* 7**,** 65-85.

MÜLLER, K. IPF within multiple domains: Generating a synthetic population for Switzerland, presentation. 11th Swiss Transport Research Conference, Ascona, 2011.

MÜLLER, K. & AXHAUSEN, K. W. 2010. *Population synthesis for microsimulation: State of the art*, ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

MÜLLER, K. & AXHAUSEN, K. W. 2012. Multi-level fitting algorithms for population synthesis. *Arbeitsberichte Verkehrs-und Raumplanung,* 821.

MURAKAMI, E. & YOUNG, J. 1997. *Daily travel by persons with low income*, US Federal Highway Administration Washington, DC.

NAGEL, K. & FLÖTTERÖD, G. Agent-based traffic assignment: Going from trips to behavioural travelers. Travel Behaviour Research in an Evolving World–Selected papers from the 12th international conference on travel behaviour research, 2012. International Association for Travel Behaviour Research, 261-294.

NAMAZI-RAD, M.-R., MOKHTARIAN, P. & PEREZ, P. 2014. Generating a dynamic synthetic population–using an age-structured two-sex model for household dynamics. *PloS one,* 9**,** e94761.

NATIONAL-RAIL. 2018. *Travel tickets on the National Rail network* [Online]. http://www.nationalrail.co.uk/times_fares/ticket_types.aspx: National-Rail Available: http://www.nationalrail.co.uk/times_fares/ticket_types.aspx [Accessed 31 May 2018 2018].

NEAL, R. M. 2001. Annealed importance sampling. *Statistics and computing,* 11**,** 125-139.

ODIARI, E. 2011. *The use of Geographic Information Systems (GIS) for Building Information Modelling (BIM).* Master of Science, CRANFIELD UNIVERSITY.

ODIARI, E. A., BIRKIN, M., GRANT-MULLER, S. AND MALLESON, N. 2016. Infilling Missing Values in Big Consumer Datasets *In:* ODIARI, E. (ed.) *CDRC - Internal Report.* University of Leeds: Leeds Institute for Data Analytics.

ODIARI, E. A., BIRKIN, M., GRANT-MULLER, S., MAGEE, T., AND MALLESON, N. 2018 (awaiting review). The use of big data in simulating attributes and mobilities of passengers on the uk rail network *Transportation Research part A: Policy and Practice*.

OGL. 2015. *How to make a freedom of information (FOI) request* [Online]. www.gov.uk/make-a-freedom-of-information-request/the-freedom-of-information-act: GOV.UK. Available: www.gov.uk/make-a-freedom-of-information-request/the-freedom-of-information-act [Accessed March 9th 2015 2015].

ONS. 2013. *Census reveals details of how we travel to work in England and Wales* [Online]. nationalarchives.gov.uk. Available: www.nationalarchives.gov.uk [Accessed 07 April 2017 2017].

ONS. 2017. *Geography data- Hierarchical Representation of UK Statistical Geographies* [Online]. ONS. Available: https://ons.maps.arcgis.com/ [Accessed November 2 2017].

OPENSHAW, S. & ALVANIDES, S. 2001. Designing zoning systems for representation of socio-economic data. *Time and Motion of Socio-Economic Units. Taylor and Francis, London*.

OPENSHAW, S. & RAO, L. 1995. Algorithms for reengineering 1991 Census geography. *Environment and planning A,* 27**,** 425-446.

ORR. 2012a. *Contract Form - Origin destination matrix 2011-12 and 2012-13* [Online]. data.gov.uk: Office of Rail Regulation. Available: data.gov.uk [Accessed 28th December 2015].

ORR. 2012b. *Origin – Destination Matrix 2011/12 Summary Report April 2013* [Online]. orr.gov.uk/: Office of Rail Regulation. Available: orr.gov.uk/ [Accessed 28th December 2015].

ORR. 2014a. *Estimates of Station Usage 2013/14* [Online]. orr.gov.uk/: Office of Rail Regulation. Available: orr.gov.uk/ [Accessed 28th December 2015].

ORR. 2014b. *Framework Agreement for Consultancy Services - Order Form* [Online]. ORR: Office of Rail Regulation. Available: data.gov.uk/ [Accessed 28th December 2015].

ORR. 2014c. *Origin - Destination Matrix 2013/14 Summary Report* [Online]. http://orr.gov.uk: Steer Davies Gleave. Available: orr.gov.uk [Accessed 28th December 2015].

ORTUZAR, J. D. D. & WILLUMSEN, L. G. 2002. *Modelling transport*.

ORTÚZAR S, J. D. D. & WILLUMSEN, L. G. 2011. *Modelling Transport,* Chichester, West Sussex, United Kingdom, John Wiley & Sons.

PADDOCK, S. M. 2002. Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika,* 89**,** 529-538.

PATRIKSSON, M. 2015. *The traffic assignment problem: models and methods*, Courier Dover Publications.

PEARL, J. 2009. *Causality*, Cambridge university press.

PEARL, J., GLYMOUR, M. & JEWELL, N. P. 2016. *Causal inference in statistics: a primer*, John Wiley & Sons.

PEEL, D. & MCLACHLAN, G. J. 2000. Robust mixture modelling using the t distribution. *Statistics and computing,* 10**,** 339-348.

PENNY, W. D., FRISTON, K. J., ASHBURNER, J. T., KIEBEL, S. J. & NICHOLS, T. E. 2011. *Statistical parametric mapping: the analysis of functional brain images*, Elsevier.

PLUMMER, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing, 2003. Vienna, Austria, 125.

PLUMMER, M., BEST, N., COWLES, K. & VINES, K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R news,* 6**,** 7-11.

PODANI, J. 1999. Extending Gower's general coefficient of similarity to ordinal characters. *Taxon***,** 331-340.

POLETTI, F., BÖSCH, P. M., CIARI, F. & AXHAUSEN, K. W. 2017. Public transit route mapping for large-scale multimodal networks. *ISPRS International Journal of Geo-Information,* 6**,** 268.

POTTHOFF, R. F., TUDOR, G. E., PIEPER, K. S. & HASSELBLAD, V. 2006. Can one assess whether missing data are missing at random in medical studies? *Statistical methods in medical research,* 15**,** 213-234.

PRESTON, J. 1987. *The evaluation of new local rail stations in West Yorkshire.* University of Leeds.

PRESTON, J. 1991. Demand forecasting for new local rail stations and services. *Journal of Transport Economics and Policy***,** 183-202.

PRESTON, J. 2012. High Speed Rail in Britain: about time or a waste of time? *Journal of Transport Geography,* 22**,** 308-311.

RAFIAH, M., RICE, J. R., FERGUSON, J. S. G., SADLER, A. J. & HARRISON, P. R. 2004. Integrated journey planner. Google Patents.

RAIL DELIVERY GROUP. 2015. *Rail Delivery Group* [Online]. www.raildeliverygroup.com: http://orr.gov.uk/about-orr/who-we-work-with/industry-organisations/rail-delivery-group. Available: http://www.raildeliverygroup.com/about-us/britains-railway-in-numbers/198-keeping-britain-on-the-move.html [Accessed 15 January 2015 2015].

RAND-CORPORATION 2017. Developing a New National Transport Model for the UK. *In:* FOX, J. (ed.). RAND corporation.

RANEY, B., CETIN, N., VÖLLMY, A., VRTIC, M., AXHAUSEN, K. & NAGEL, K. 2003. An agent-based microsimulation model of Swiss travel: First results. *Networks and Spatial Economics,* 3**,** 23-41.

RAO, S. S. & RAO, S. S. 2009. *Engineering optimization: theory and practice*, John Wiley & Sons.

RAUDENBUSH, S. W. & BRYK, A. S. 2002. *Hierarchical linear models: Applications and data analysis methods*, Sage.

RAYPORT, J. F. 2011. What Big Data Needs: A Code of Ethical Practices. *MIT Technology Review* Cambridge, Massachusetts Jason Pontin.

REES, P. 1994. Estimating and projecting the populations of urban communities. *Environment & planning A,* 26**,** 1,671-97.

RHOADS, C. H. 2012. Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy,* 3.

RICHARDSON, A. J., AMPT, E. S. & MEYBURG, A. H. 1995. *Survey methods for transport planning*, Eucalyptus Press Melbourne.

RICHARDSON, S. & BEST, N. 2003. Bayesian hierarchical models in ecological studies of health–environment effects. *Environmetrics,* 14**,** 129-147.

RICHEY, M. 2010. The evolution of Markov chain Monte Carlo methods. *American Mathematical Monthly,* 117**,** 383-413.

RIESER, M., NAGEL, K., BEUCK, U., BALMER, M. & RÜMENAPP, J. 2007. Agent-oriented coupling of activity-based demand generation with multiagent traffic simulation. *Transportation Research Record: Journal of the Transportation Research Board***,** 10-17.

RIPLEY, B. D. 2005. *Spatial statistics*, John Wiley & Sons.

ROBERT, C. P. 2004. *Monte carlo methods*, Wiley Online Library.

ROBERTS, G. O. & ROSENTHAL, J. S. 2004. General state space Markov chains and MCMC algorithms. *Probability surveys,* 1**,** 20-71.

RODRıGUEZ, G. 2013. Models for count data with overdispersion.

ROEDER, K. 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association,* 85**,** 617-624.

ROSENTHAL, J. S. 2011. Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo***,** 93-112.

ROY, J., LUM, K. J., DANIELS, M. J., ZELDOW, B., DWORKIN, J. & RE III, V. L. 2017. Bayesian nonparametric generative models for causal inference with missing at random covariates. *arXiv preprint arXiv:1702.08496*.

ROYSTON, P. 2005. Multiple imputation of missing values: update of ice. *Stata Journal,* 5**,** 527.

RSTUDIOTEAM 2016. RStudio: Integrated Development Environment for R. RStudio, Inc.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika,* 63**,** 581-592.

RUBIN, D. B. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. Proceedings of the survey research methods section of the American Statistical Association, 1978. American Statistical Association, 20-34.

RUBIN, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American statistical Association,* 91**,** 473-489.

RUBIN, D. B. 2004. *Multiple imputation for nonresponse in surveys*, John Wiley & Sons.

RUBIN, D. B. & SCHENKER, N. 1991. Multiple imputation in health-are databases: An overview and some applications. *Statistics in medicine,* 10**,** 585-598.

RUSCHENDORF, L. 1995. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics,* 23**,** 1160-1174.

SALVINI, P. & MILLER, E. J. 2005. ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and Spatial Economics,* 5**,** 217-234.

SCHAFER, J. L. 1997. *Analysis of incomplete multivariate data*, CRC press.

SCHAFER, J. L. & GRAHAM, J. W. 2002. Missing data: our view of the state of the art. *Psychological methods,* 7**,** 147.

SDG 2011. Research on Ticketing in Major Urban Areas. *In:* SDG (ed.) *Research on Ticketing in Major Urban Areas.* Passenger Demand Forecasting Council (PDFC).

SDG 2012. Demand in the Centro PTE area. *Research on PTE Tickets* Passeneger Demand Forecasting Council - ATOC.

SDG 2014. Origin-Destination Matrix 2013/14 Summary Report. *In:* GLEAVE, S. D. (ed.) *Origin-Destination Matrix.* www://orr.gov.uk: Office of Rail and Road.

SDG 2016. Origin-Destination Matrix 2014/15 Summary Report. *In:* GLEAVE, S. D. (ed.) *Origin-Destination Matrix.* www://orr.gov.uk: Office of Rail and Road.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*, Wadsworth Cengage learning.

SHAH, A. D., BARTLETT, J. W., CARPENTER, J., NICHOLAS, O. & HEMINGWAY, H. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *American Journal of Epidemiology,* 179**,** 764-774.

SHEPARD, D. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 23rd ACM national conference, 1968. ACM, 517-524.

SHORE, J. & JOHNSON, R. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory,* 26**,** 26-37.

SIDDIQUI, O. & ALI, M. W. 1998. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of biopharmaceutical statistics,* 8**,** 545-563.

SIMINI, F., GONZÁLEZ, M. C., MARITAN, A. & BARABÁSI, A.-L. 2012. A universal model for mobility and migration patterns. *Nature,* 484**,** 96-100.

SMITH, A. F. & ROBERTS, G. O. 1993. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3-23.

SMITH, D. M., CLARKE, G. P. & HARLAND, K. 2009. Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A,* 41**,** 1251-1268.

SOLLACI, L. B. & PEREIRA, M. G. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the medical library association,* 92**,** 364.

SOLTOW, L. 1960. The distribution of income related to changes in the distributions of education, age, and occupation. *The Review of Economics and Statistics***,** 450-453.

SPIEGELHALTER, D., THOMAS, A., BEST, N. & LUNN, D. 2007. OpenBUGS user manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge*.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. & VAN DER LINDE, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 64**,** 583-639.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. G., GILKS, W. & LUNN, D. 1996. BUGS: Bayesian inference using Gibbs sampling. *Version 0.5,(version ii) http://www. mrc-bsu. cam. ac. uk/bugs,* 19.

SPIESS, H. & FLORIAN, M. 1989. Optimal strategies: a new assignment model for transit networks. *Transportation Research Part B: Methodological,* 23**,** 83-102.

STEIN, M. L. 2012. *Interpolation of spatial data: some theory for kriging*, Springer Science & Business Media.

STEKHOVEN, D. J. & BÜHLMANN, P. 2011. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics,* 28**,** 112-118.

STEKHOVEN, D. J. & BÜHLMANN, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics,* 28**,** 112-118.

STERNE, J. A., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. & CARPENTER, J. R. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj,* 338**,** b2393.

STILLWELL, J. 1978. Interzonal migration: some historical tests of spatial-interaction models. *Environment and Planning A,* 10**,** 1187-1200.

STILLWELL, J. 2006. Providing access to census-based interaction data in the UK: that's WICID. *The Journal of Systemics, Cybernetics and Informatics,* 4**,** 63-68.

STILLWELL, J. & DUKE-WILLIAMS, O. 2003. A new web-based interface to British census of population origin–destination statistics. *Environment and Planning A,* 35**,** 113-132.

STURTZ, S., LIGGES, U. & GELMAN, A. 2010. R2OpenBUGS: a package for running OpenBUGS from R. *URL http://cran. rproject. org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS. pdf*.

SUN, L. & ERATH, A. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies,* 61**,** 49-62.

SUN, W. & YUAN, Y.-X. 2006. *Optimization theory and methods: nonlinear programming*, Springer Science & Business Media.

SUZUKI, J. 2005. The lost calculus (1637-1670): Tangency and optimization without limits. *Mathematics Magazine,* 78**,** 339-353.

TANTON, R. 2014. A review of spatial microsimulation methods. *International Journal of Microsimulation,* 7**,** 4-25.

TANTON, R. & EDWARDS, K. 2012. *Spatial microsimulation: A reference guide for users*, Springer Science & Business Media.

TAYLOR, J. 2013a. Robust Foundations. *In:* FICKLING, R. (ed.) *Rail in the North Demand and Revenue Model.* \\UKMANCFP02\Projects(Odd)\313427 Rail in the North\4.0 Reporting\Base Matrix Report: Mott MacDonald.

TAYLOR, J. 2013b. Robust Foundations - Rail in the North Demand and Revenue Model. *Rail in the North Demand and Revenue Model - Base Matrix Report.* Mott MacDonald, University of Leeds, and MVA Consultancy.

TEAM, R. C. 2000. R language definition. *Vienna, Austria: R foundation for statistical computing*.

TENNANT, P. 8th September 2018 2018. *RE: Causal inference with observational data: challenges and pitfalls.* Type to ODIARI, E.

THERNEAU, T., ATKINSON, B., RIPLEY, B. & RIPLEY, M. B. 2017. Package ‘rpart’. *Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016)*.

TÖRN, A. & ZILINSKAS, A. 1989. *Global optimization*, Springer.

TRAIN, K. E., MCFADDEN, D. L. & BEN-AKIVA, M. 1987. The demand for local telephone service: A fully discrete model of residential calling patterns and service choices. *The RAND Journal of Economics***,** 109-123.

TRANSITFEEDS. 2017. *ATOC GTFS TransitFeeds* [Online]. www.transitfeeds.com/: TransitFeeds. Available: https://transitfeeds.com/ [Accessed July 26 2017].

TREASURY, H. M. S. 2011. National infrastructure plan 2011. *Infrastructure UK, UK Treasury Department (HM Treasury), http://cdn. hm-treasury. gov. uk/national_infrastructure_plan291111. pdf*.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. & ALTMAN, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics,* 17**,** 520-525.

TU, Y.-K., KELLETT, M., CLEREHUGH, V. & GILTHORPE, M. S. 2005. Problems of correlations between explanatory variables in multiple regression analyses in the dental literature. *British dental journal,* 199**,** 457-461.

TURNBULL, B. W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 290-295.

UK-DATA-SERVICE 2011. 2011 Census United Kingdom - Safeguarded. 08-12-2014 ed.: ONS

UK-LEGISLATION. 2018. *Data Protection Act 2018* [Online]. Available: http://www.legislation.gov.uk/ [Accessed 31 May 2018 2018].

UKDS 2016. Census Support: Flow Data. *Flow Data* 15-01-2015 ed. www.wicid.ukdataservice.ac.uk/: ONS / SASPAC

UNIVERSITY-OF-LEEDS. 2017. *Advanced Research Computing* [Online]. University of Leeds. Available: http://arc.leeds.ac.uk/systems/arc3/ [Accessed 15 April 2017 2017].

UOL. 2015. *Good research practice & research ethics* [Online]. University of Leeds. Available: www.leeds.ac.uk/ethics [Accessed 5th June 2015 2015].

UPTON, G. J. 1985. Modelling cross-tabulated regional data. *Nijkamp P, Leitner H, Wrigley, N Martinus, Measuring the Unmeasurable***,** 197-218.

VAN BUUREN, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research,* 16**,** 219-242.

VAN BUUREN, S., BOSHUIZEN, H. C. & KNOOK, D. L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine,* 18**,** 681-694.

VAN DER HEIJDEN, G. J., DONDERS, A. R. T., STIJNEN, T. & MOONS, K. G. 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *Journal of clinical epidemiology,* 59**,** 1102-1109.

VAN NES, R. 2002. Design of multimodal transport networks: A hierarchical approach.

VARIAN, H. R. 2014. Big data: New tricks for econometrics. *The Journal of Economic Perspectives,* 28**,** 3-27.

VICCARS, A. 2002. *Report to the Executive for Decision 4 November 2002* [Online]. Fareham Borough Council. Available: http://www.fareham.gov.uk/crs/executive/021104/reports-public/xpt-021104-r18-avi.pdf Strategic Rail Authority Consultation - Future Fares Policy].

VIDYATTAMA, Y., MIRANTI, R., MCNAMARA, J., TANTON, R. & HARDING, A. 2011. Issues in spatial microsimulation estimation: a case study of child poverty.

VIKTOR, M.-S. & KENNETH, C. 2013. Big data: A revolution that will transform how we live, work, and think. *Houghton Mifflin Harcourt*.

WARDMAN, M., LYTHGOE, W. & WHELAN, G. 2007. Rail passenger demand forecasting: cross-sectional models revisited. *Research in transportation economics,* 20**,** 119-152.

WEBER, G. M., MANDL, K. D. & KOHANE, I. S. 2014. Finding the missing link for big biomedical data. *Jama,* 311**,** 2479-2480.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine,* 30**,** 377-399.

WHITEING, T., BROWNE, M. & ALLEN, J. 2003. City logistics: the continuing search for sustainable solutions. *Global Logistics and Distribution Planning,* 4**,** 308-332.

WHITTAKER, J. M. 1935. *Interpolatory function theory*, The University Press.

WILLETTS, D. 2013. *Eight great technologies*, Policy Exchange.

WILLIAMSON, P., BIRKIN, M. & REES, P. The simulation of whole populations using data from small area statistics, samples of anonymised records and national surveys. Research on 1991 Census conference', University of Newcastle, September, 1993.

WILLIAMSON, P., BIRKIN, M. & REES, P. H. 1998. The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A,* 30**,** 785-816.

WILLUMSEN, L. 1981. Simplified transport models based on traffic counts. *Transportation,* 10**,** 257-278.

WILLUMSEN, L. G. 1993. MULTI-MODAL MODELLING IN CONGESTED NETWORKS: SATURN AND SATCHMO. *Traffic engineering & control*.

WILSON, A. 2010. Entropy in Urban and Regional Modelling: Retrospect and Prospect. 城市和区域建模中的熵： 回顾与展望. *Geographical Analysis,* 42**,** 364-394.

WILSON, A. G. 1969. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy***,** 108-126.

WILSON, A. G. 1971. A family of spatial interaction models, and associated developments. *Environment and Planning A,* 3**,** 1-32.

WONG, D. W. 1992. The Reliability of Using the Iterative Proportional Fitting Procedure∗. *The Professional Geographer,* 44**,** 340-348.

WORSLEY, T. 2012. Rail Demand Forecasting Using the Passenger Demand Forecasting Handbook. *On the Move–Supporting Paper,* 2**,** 89-91.

WRIGHT, E. O. & PERRONE, L. 1977. Marxist class categories and income inequality. *American Sociological Review***,** 32-55.

WYCA 2016. West Yorkshire Transport Strategy Evidence Base. *Transport Strategy 2016-2036.* www.westyorks-ca.gov.uk/: West Yorkshire Combined Authority

WYCA. 2018. *M-Metro Transport for West Yorkshire* [Online]. Transport for West Yorkshire. Available: https://www.wymetro.com/trains/ [Accessed 31 May 2018 2018].

YUAN, K.-H. & BENTLER, P. M. 2000. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology,* 30**,** 165-200.

YUAN, Y. C. 2010. Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD,* 49**,** 1-11.

ZEILEIS, A., KLEIBER, C. & JACKMAN, S. 2008. Regression models for count data in R. *Journal of statistical software,* 27**,** 1-25.

ZHANG, L., XIONG, C. & BERGER, K. Multimodal Inter-Regional Origin-Destination Demand Estimation.

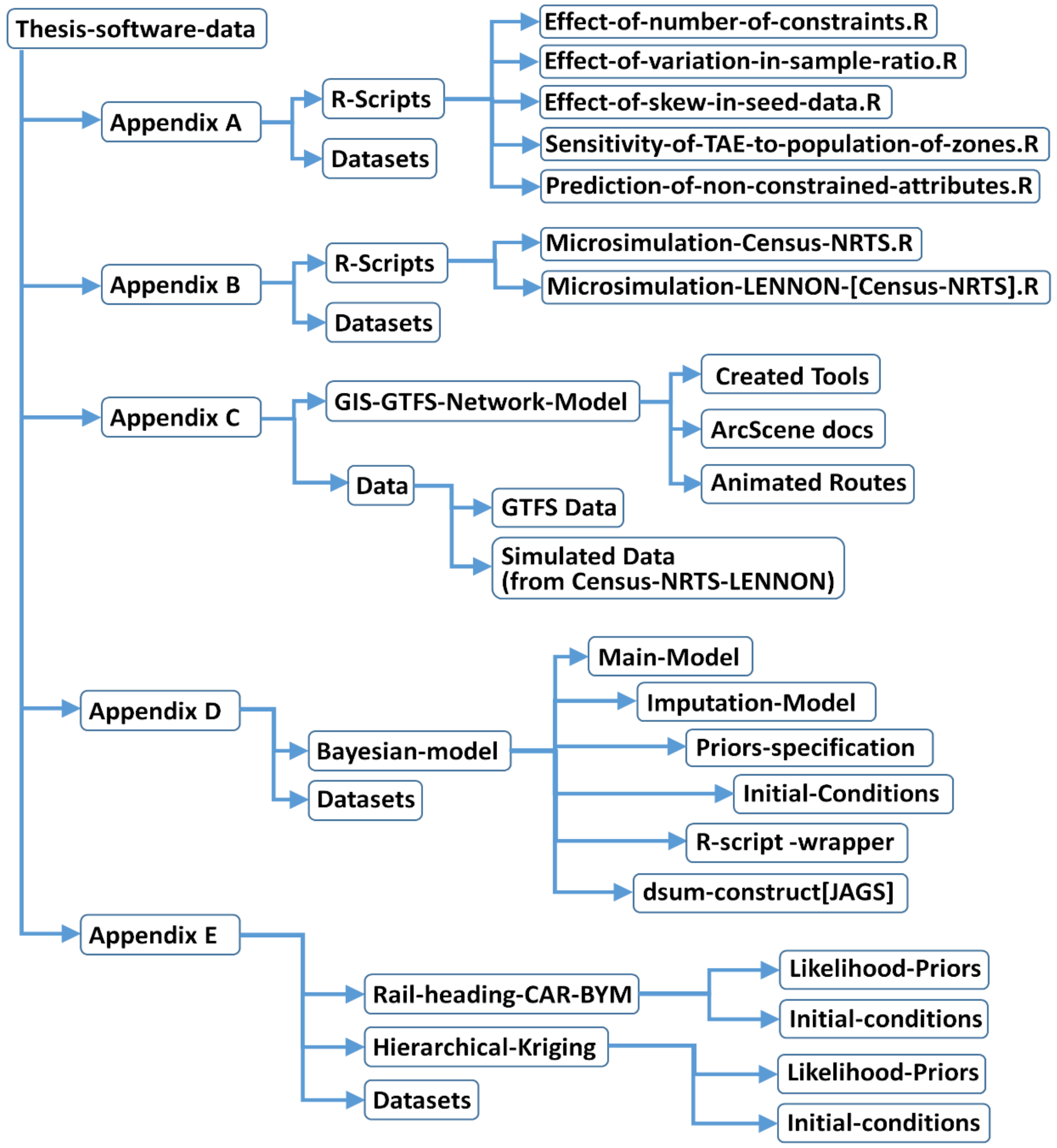
ZHAO, Y. & KOCKELMAN, K. M. 2002. The propagation of uncertainty through travel demand models: an exploratory analysis. *The Annals of regional science,* 36**,** 145-163.

ZIPF, G. K. 1946. The P 1 P 2/D hypothesis: on the intercity movement of persons. *American sociological review,* 11**,** 677-686.

Appendices

This section summarises the directories and file structure within the compact disc (CD) which accompanies this thesis. The CD contains all the software codes referred to in the thesis, and 'readme' files. Such information provided enables the location of the R-software scripts, GIS-GTFS models, WinBUGS, OpenBUGS, and JAGS scripts developed as part of the research. The CD provided was deemed necessary to enable a reproduction of the research results and findings and to enable future development of the project by other practitioners. For those datasets not available for public use, the schemata are included as this enables the creation of proxy data.

Structure of CD directory - CD:\\Thesis-software-data



Appendix A
R-script for spatial microsimulation experiment

The pertinent aspects of the codes are the checks to ensure that the list of variables and variable-categories are the same for the seed and the target datasets. This similarity in structure of the datasets forms the basis for reconciling them, also ensuring that the seed would be optimally replicated. The algorithms search all the variable-categories in the target and seed data to make sure that each variable category is consistent between the target and seed. By so doing, the micro-population created references the correct geographic zone, variable, and variable-category. This ensures the quality of the simulated population.

The R-script codes within this Appendix A are aimed at assessing the behaviour of standard spatial microsimulation methods when the target and seed data have features that are not typical of randomized control data from traditional measured stated surveys. The data used to implement the R-script codes listed in Appendix A are derived by post-processing and sampling the 2004/5 NRTS dataset.

A.1 Effect of number of constraints

This is an R-script which assesses if an increase in number of constraint variables used in spatial microsimulation results in lower values of total absolute error (TAE). This algorithm is implemented for both the deterministic and the stochastic strategies.

A.2 Effect of variations in sample ratios

The R-script included here assesses the effect of various sample ratios of the seed data on the results of the deterministic Iterative Proportional Fitting (IPF), the stochastic Hill Climbing (HC), and the Simulated Annealing (SA) spatial microsimulation methodologies. In effect, this experiment investigates the deterministic and stochastic simulation methodologies and assesses how different seed sample ratios would affect the accuracy of a simulated population. A Monte Carlo sampling is included in the spatial microsimulation methodology to attenuate out the effects of choice of particular variable values in a seed of specific sample ratio.

A.3 Effect of skew in seed data

The script evaluates the ability of spatial microsimulation methods when the seed data is skewed relative to the target distribution. This is an important aspect of the research as a typical feature of consumer data is that it tends to be a sewed subset of the wider population. The established methods for optimisation typically used to reconcile disparate datasets all assume that the datasets being combined have similar distributions. The distribution of railway ticketing data has about 40% rail commuters, whilst the wider the distribution of the wider Census population has about 8% rail commuters. As such, the distribution of the rail ticketing data is skewed relative to the wider population. Observational data revealed from consumers of digital services tend to be skewed relative to traditional analysis data from randomized control surveys. As a result, these novel consumer data need to be validated prior to combining with traditional survey data. This experiment ascertains the behaviour of deterministic and stochastic microsimulation methods when the seed data is designed to be skewed relative to the target.

A.4 Sensitivity of TAE to population of zones

This script is intended to assess the ability of the spatial microsimulation methods in yielding accurate simulated populations for zones that are relatively small compared to other zones. This is achieved by adapting the code in Appendix A.1 and assessing how the TAE values change for peripheral Postcode Areas of lower population (like OL, S, and HX), when compared to larger Postcode Areas of higher population (like LS, BD, and WF). Note that the population of some of the Postcode Areas are low simply because the full Area was not taken into consideration as it was clipped to fit into the West Yorkshire study area.

A.5 Prediction of non-constrained attributes

This script assesses the ability of the spatial microsimulation methods to predict those variables within the target, but which were not included as constraints in the spatial microsimulation process. This is achieved by adapting the script in Appendix A.3, by running the model with fewer than 8 constraint variables as appropriate, and then estimating the difference in TAE for those variables which were not included as target constraints in the spatial microsimulation process.

Appendix B
R-script for stage 1 and stage 2 spatial microsimulation

The R-scripts included in Appendix B are used to combine the Census and NRTS datasets, and then further to combines these with the LENNON ticketing data. The scripts developed have been adapted to accommodate situations where the seed is skewed relative to the target. This is achieved by attaching an origin and destination to each individual's socio-demographic attribute. Using such a dataset ensured that the seed is fit to the full conditional distribution of the target. This facilitates a 3-level hierarchical constraint, which in-turn ensures that not only are better TAE values achieved, but the seed also yields the distribution of the target.

In practical implementations, it was necessary to augment the seed data by addition of a nominal few seed data of requisite characteristics to ensure consistency with the aggregate target, while monitoring their emergence in the final simulated population. On other occasions it has been prudent to reduce the target volumes to be commensurate with the seed, but caution is necessary in doing so, to ensure that the errors or discrepancies introduced does not affect the data structure. A further precaution is to ensure that in the standard IPF implementation the marginal across each aggregate constraint are equal, as the variables in the aggregate table represent the same population. There are also possibilities of opting for probabilities such that the marginal are normalised to sum to one

B.1 Stage 1 R-script (Census-NRTS)

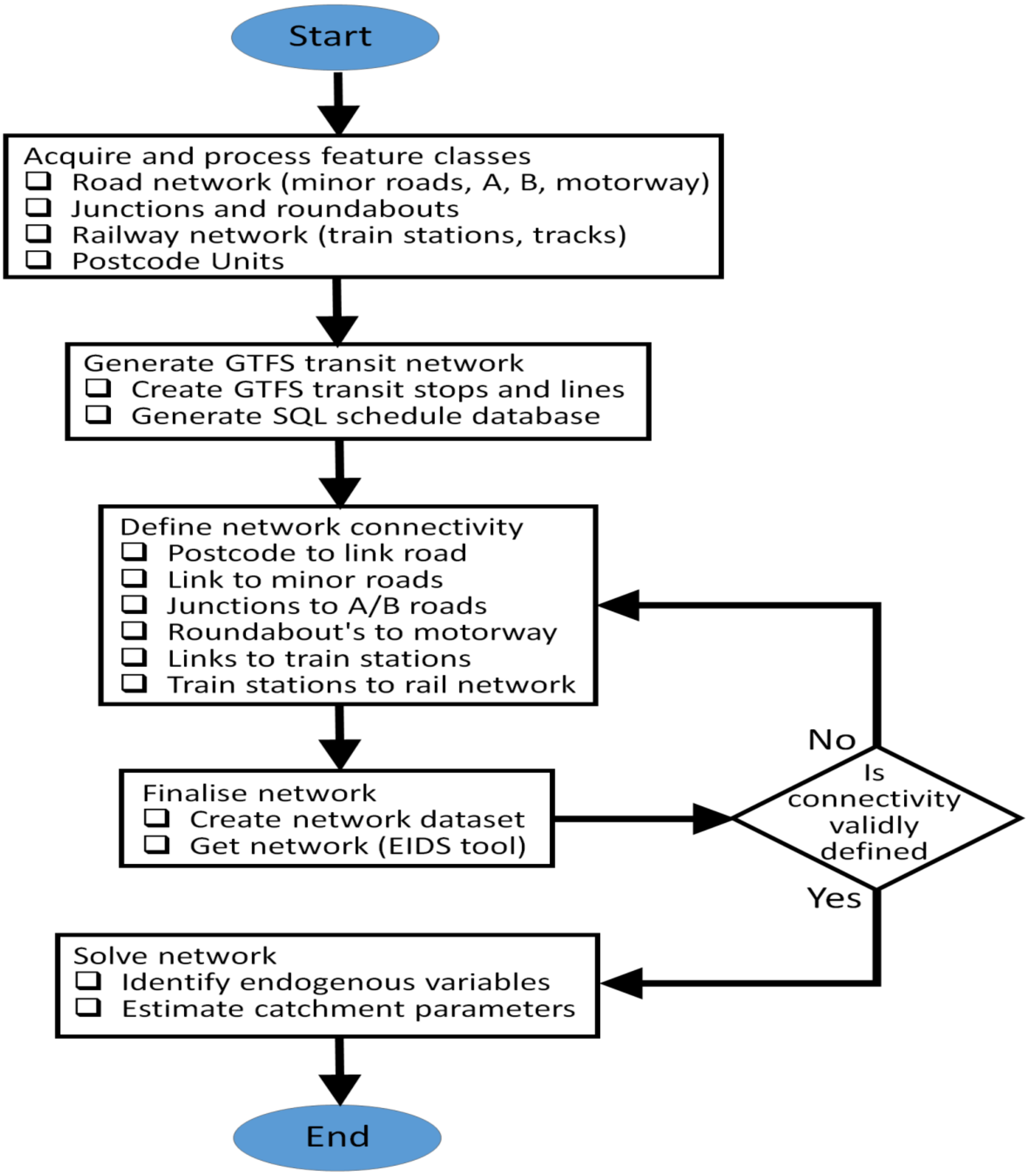
The R-script developed to implement the Stage 1 spatial microsimulation combines the 2011 Census interaction data (flow data) with the 2004/5 National Rail Travel Survey (NRTS) datasets. The m-IPF spatial microsimulation methodology is developed to facilitate a 3-level simultaneous hierarchical constraint. Note that the NRTS is a skewed sub-set of the Census population.

B.2 Stage 2 R-script (Census-NRTS-LENNON)

Stage 2 is the spatial microsimulation conducted to combine the simulated population from Stage 1 with the LENNON ticketing data. The LENNON is used as the target dataset.

Appendix C
ArcGIS creation of GIS-GTFS network model

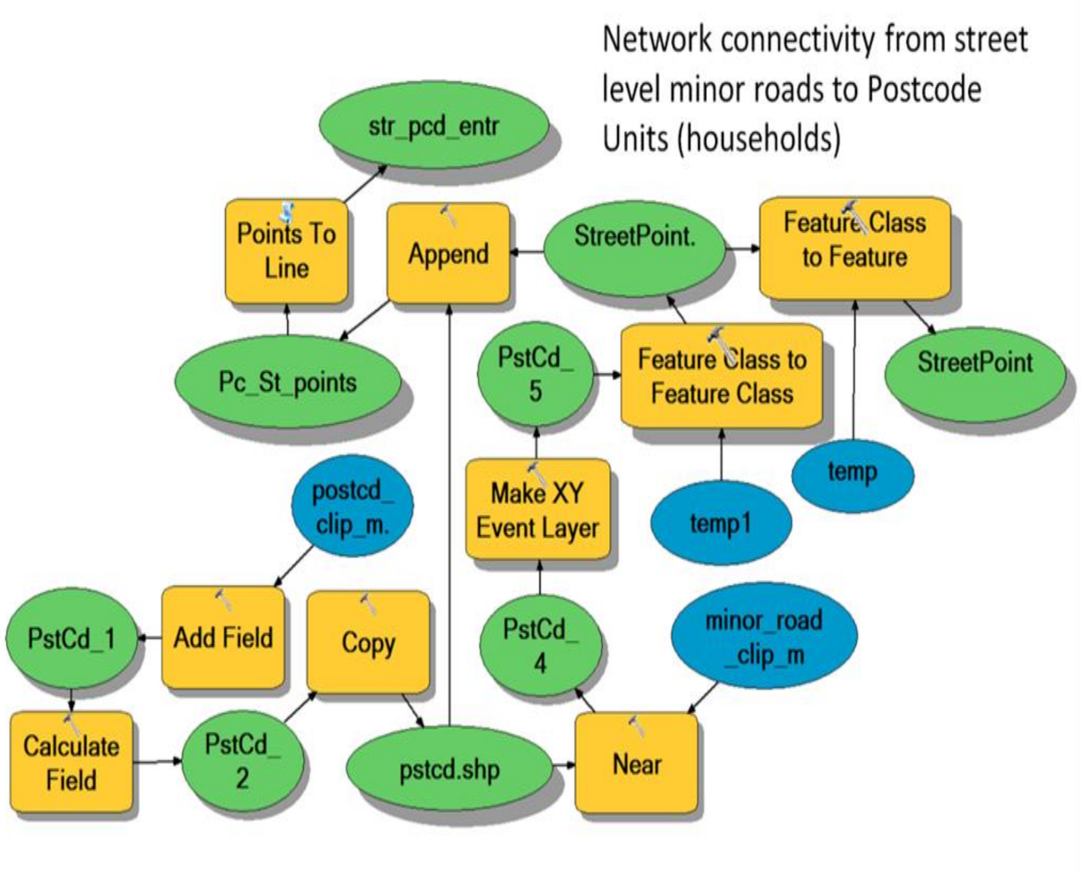
A range of ArcGIS models have been created to facilitate developing the GIS-GTFS network model. The model builder software has been chosen to enable the stages adopted in creating aspects of the GIS-GTFS model to be visualised. The stages are summarized in the flow chart below. The network dataset created needed quality assessment to ensure connectivity at peripheral zones where adequate splits on features might not have been successful created due to long distances between Postcode Units and minor roads. The detailed ArcGIS Model builder and the datasets are included in the CD accompanying the thesis as described at the beginning of the Appendices.

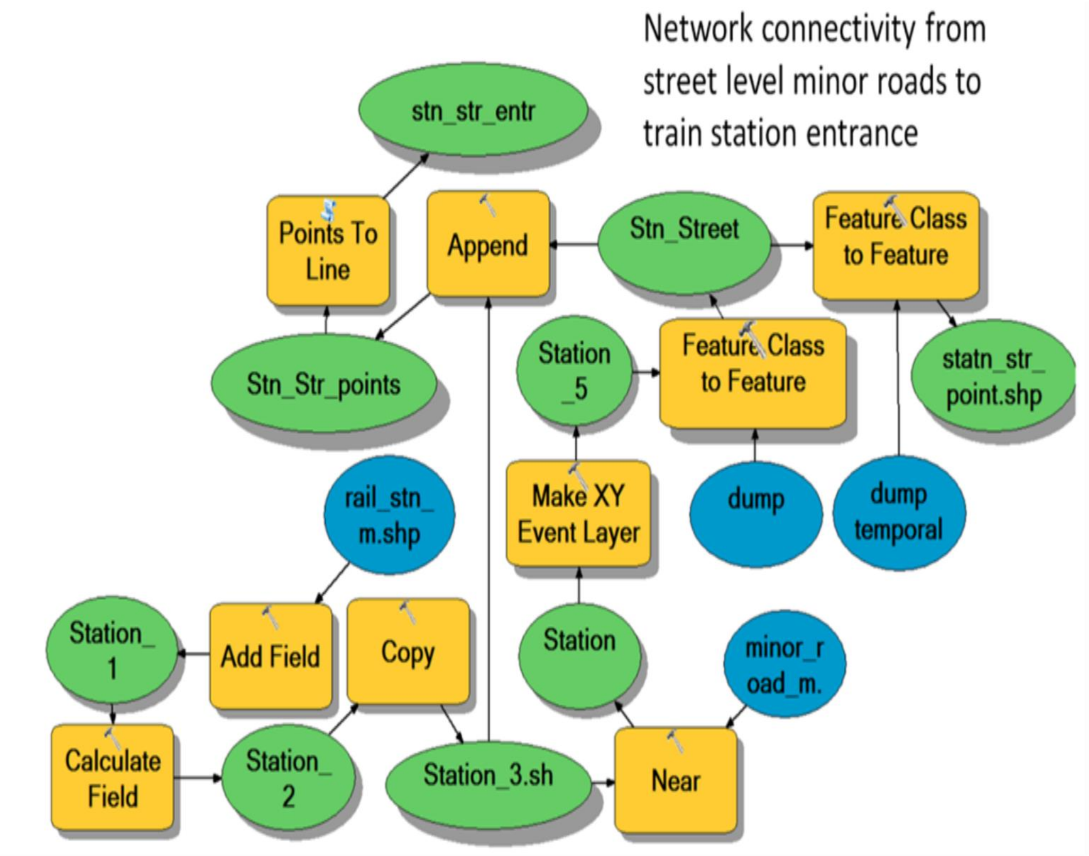

Start
Acquire and process feature classes
❑ Road network (minor roads, A, B, motorway)
❑ Junctions and roundabouts
❑ Railway network (train stations, tracks)
❑ Postcode Units
Generate GTFS transit network
❑ Create GTFS transit stops and lines
❑ Generate SQL schedule database
Define network connectivity
❑ Postcode to link road
❑ Link to minor roads
❑ Junctions to A/B roads
❑ Roundabout's to motorway
❑ Links to train stations
❑ Train stations to rail network
Finalise network
❑ Create network dataset
❑ Get network (EIDS tool)
No
Is connectivity validly defined
Yes
Solve network
❑ Identify endogenous variables
❑ Estimate catchment parameters
End

C.1 Model used to create network linkages between Postcode Units, road networks, and the railways

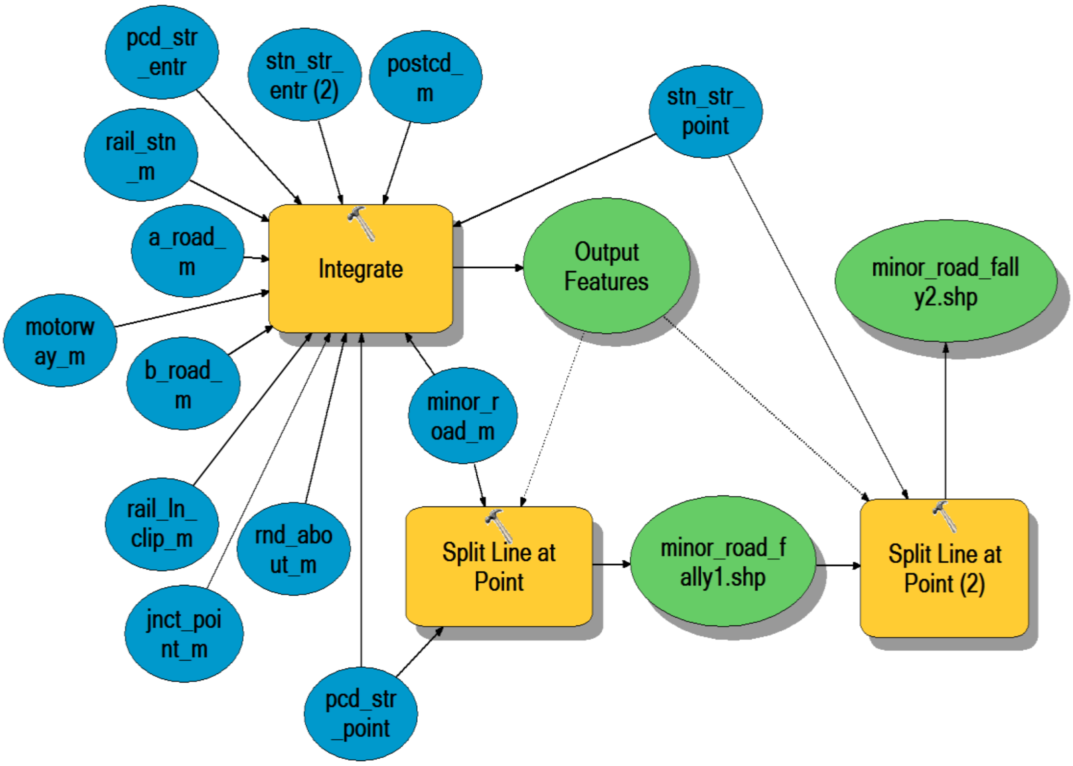
The ArcGIS model builder created to build logistical network links between the Postcode Units, road entrances, minor roads, A and B roads, motorways, railways entrances, railway stations, and railway trains and railway lines are shown here. The details are included in the CD accompanying the thesis

Two sets of line features respectively connecting the Postcode Units to the minor roads, and the train stations to the road network are created, such that there are no topological conflicts in the GIS network created. The connectors make provision for where the passengers can access the transit lines and the Postcode Units. The connectors also enable modelling of delays in boarding the train, as well as restrict access to particular train stations for those passengers with disability, when there are no disable provisions at a particular train station. A precaution in building the connecting links is to establish functional station from the GTFS database as otherwise there is a risk of including connections for the many disused stations (e.g. Milnrow, New Hey, Derker etc.) previously on the Lancashire and Yorkshire Railways which was closed in 2009), but still exist on conventional databases of UK railway stations.



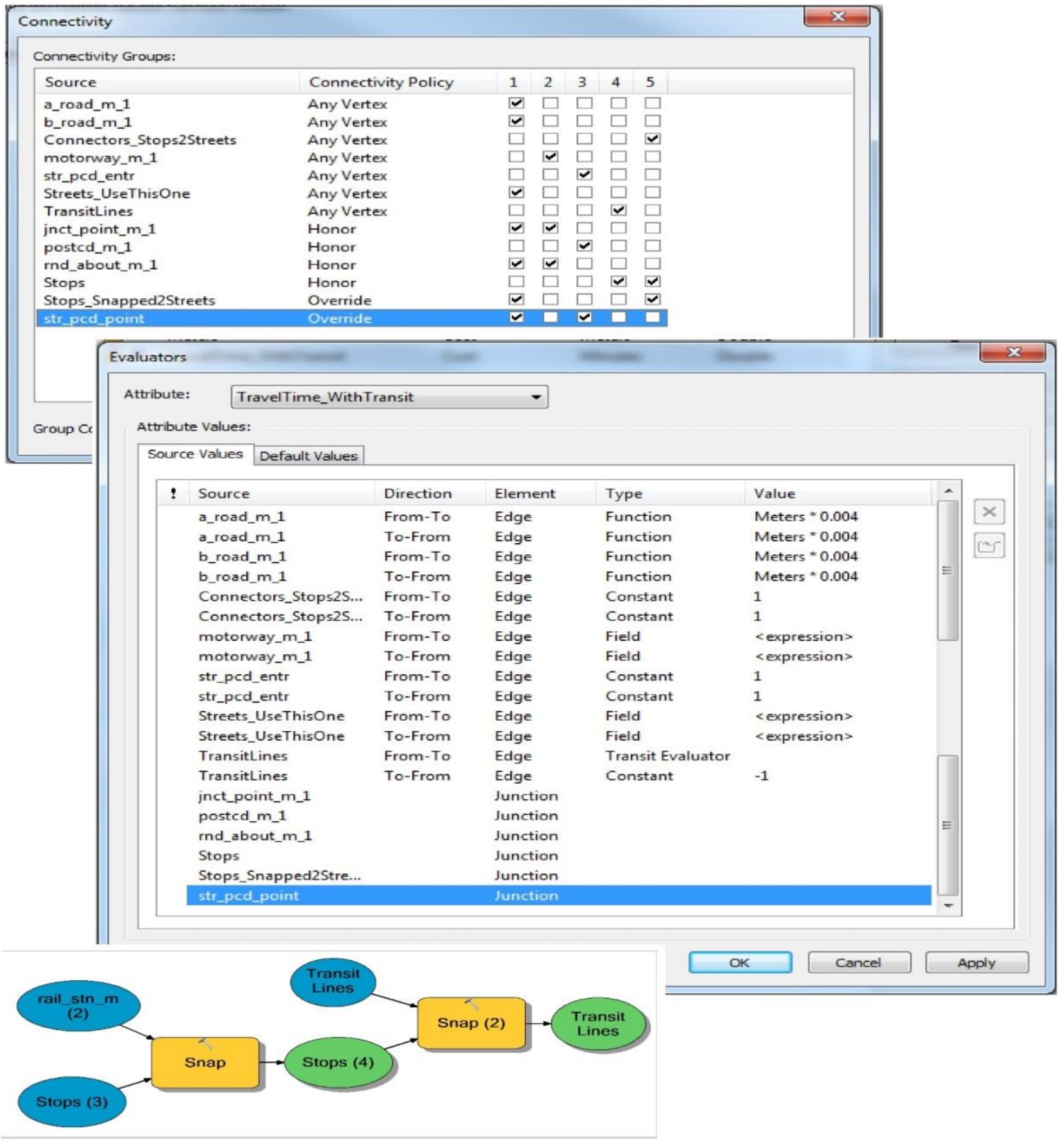


The integrity of shared feature vertexes are validated, and connectivity points between Postcode Units and minor roads, and between railway stations and road network are created by splitting the features appropriately. This integration is achieved using the model builder below.



C.2 Detailed specification of the connectivity between elements of the transport network

The connectivity rules between elements of the transport network define the manner in which the network is traversed. As detailed in section 6.2.2., the connectivity ensures that for instance it is not possible to travel from the train network to the roads other than through the railway stations, station entry points and non-motorway roads. In the figure below, connectivity between the feature classes are defined in the top table. The GTFS SQL database enables query of transit schedule during network analysis. The travel time evaluator within ArcGIS calculates the traversal time when solving Network Analyst problems. The embedded ArcGIS model is crucial in matching (snapping) the GTFS transit stops and transit ends with the station and track shapefiles from Ordnance Survey.



Connectivity

Connectivity Groups:

Source	Connectivity Policy	1	2	3	4	5
a_road_m_1	Any Vertex	☑	□	□	□	□
b_road_m_1	Any Vertex	☑	□	□	□	□
Connectors_Stops2Streets	Any Vertex	□	□	□	□	☑
motorway_m_1	Any Vertex	□	☑	□	□	□
str_pcd_entr	Any Vertex	□	□	☑	□	□
Streets_UseThisOne	Any Vertex	☑	□	□	□	□
TransitLines	Any Vertex	□	□	□	☑	□
jnct_point_m_1	Honor	☑	☑	□	□	□
postcd_m_1	Honor	□	□	☑	□	□
rnd_about_m_1	Honor	☑	☑	□	□	□
Stops	Honor	□	□	□	☑	☑
Stops_Snapped2Streets	Override	☑	□	□	□	☑
str_pcd_point	Override	☑	□	☑	□	□

Group Cc

Evaluators

Attribute: TravelTime_WithTransit

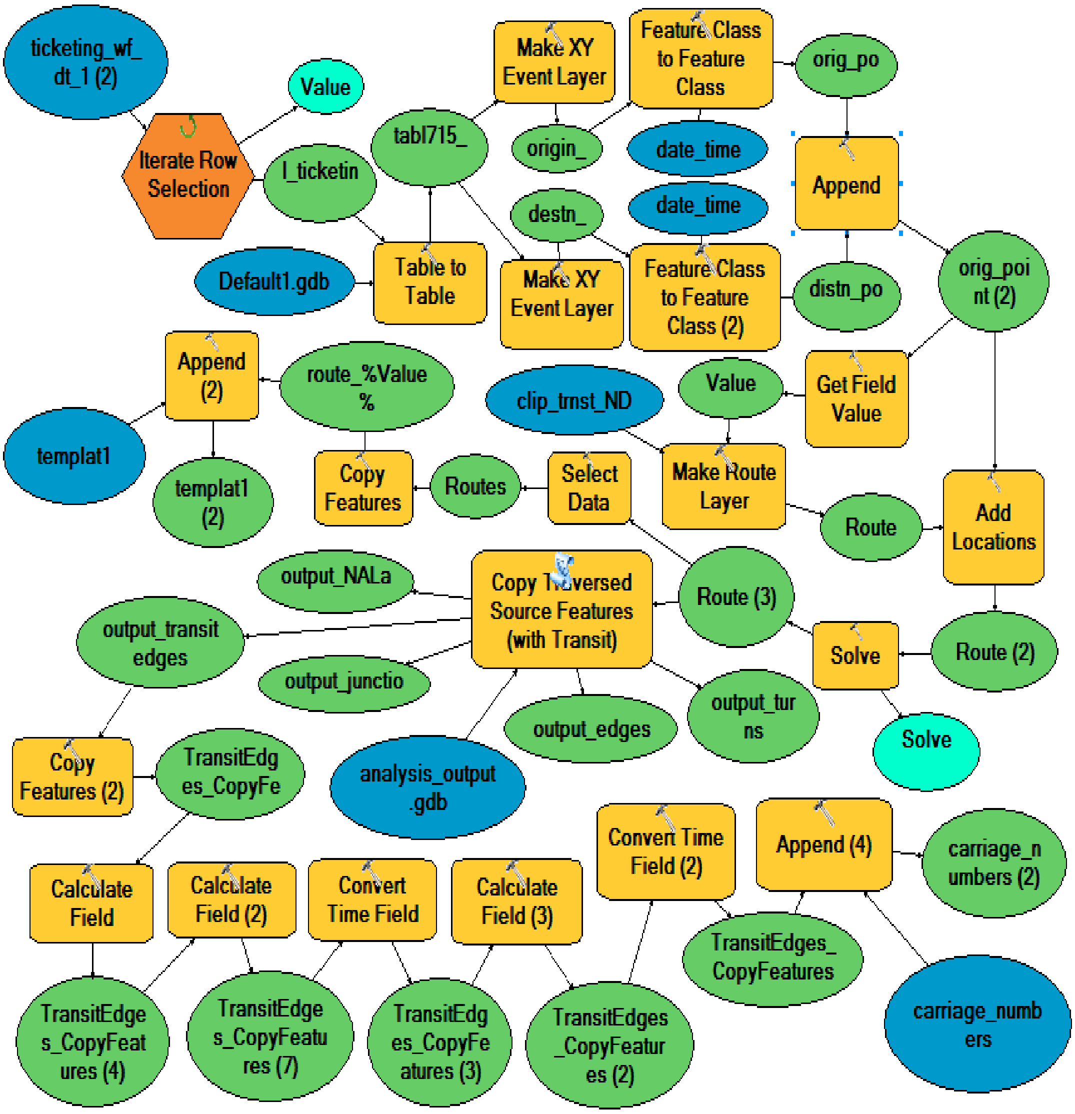
Attribute Values:

Source Values | Default Values

!	Source	Direction	Element	Type	Value
	a_road_m_1	From-To	Edge	Function	Meters * 0.004
	a_road_m_1	To-From	Edge	Function	Meters * 0.004
	b_road_m_1	From-To	Edge	Function	Meters * 0.004
	b_road_m_1	To-From	Edge	Function	Meters * 0.004
	Connectors_Stops2S...	From-To	Edge	Constant	1
	Connectors_Stops2S...	To-From	Edge	Constant	1
	motorway_m_1	From-To	Edge	Field	<expression>
	motorway_m_1	To-From	Edge	Field	<expression>
	str_pcd_entr	From-To	Edge	Constant	1
	str_pcd_entr	To-From	Edge	Constant	1
	Streets_UseThisOne	From-To	Edge	Field	<expression>
	Streets_UseThisOne	To-From	Edge	Field	<expression>
	TransitLines	From-To	Edge	Transit Evaluator	
	TransitLines	To-From	Edge	Constant	-1
	jnct_point_m_1		Junction		
	postcd_m_1		Junction		
	rnd_about_m_1		Junction		
	Stops		Junction		
	Stops_Snapped2Stre...		Junction		
	str_pcd_point		Junction		

C.3 Model used to iteratively solve the detailed mobility of each passenger on the rail network

The ArcGIS model builder created to solve the detailed mobility of each passenger on the rail network. This model was created to identify the additional endogenous passenger attributes like waiting times, number of stops, average speed of trains, passenger access to rail service provision, actual distance travelled, etc. This ArcGIS model is adapted for use to estimate passenger volumes in the trains, as well as to facilitate animation of transit trains with passengers embedded. Each of the simulated passengers has a time of arrival at station, time of first train, specifications of intermediate stations. This information assists in forming the basis for solving each passenger's mobility. The route and times for trains are derived from the GTFS information. Further details of these ArcGIS models are included in the CD accompanying the thesis.



Appendix D
Bayesian model for imputation and mobility analysis

In Appendix D, the Bayesian models developed in the WinBUGS, OpenBUGS, and JAGS environments are specified in more detail. Included are the specifications of the data likelihood functions, the system parameter priors, and initial conditions. The Bayesian models also include a specification of missing data models in sections where these were developed. In the JAGS implementation of the Bayesian models, additional separate models are created to distinguish the procedural R-Software traditional statistical modelling platform from the declarative BUGS implementation of the Bayesian models.

The main model is included in Section D.1, the missing data imputation models are included in Section D.2, the specification of the priors are included in Section D.3, the initial conditions are included in Section D.4, the R-script code that wraps all the components of the Bayesian model is included in Section D.5. The detailed R-scripts developed to implement the Bayesian modelling are included in the CD accompanying the thesis.

D.1 Main model

The main mobility model has negative Binomial data likelihood, and the continuous parameters are centred. The mobility spatial-interaction is accounted for by the inclusion of the categorical origins and destinations. Non-linearity is introduced into the model by the inclusion of a distance (log-distance) decay representing the propensity of passengers to resist longer travel distances. The missing journey factors and ticket validity periods are treated as hierarchical phenomena.

D.2 Missing data imputation model

In the Bayesian framework, the missing data are treated as random variables, and additional covariate models are built using the partially observed variables (those with missing data values) as the outcome variable. The outcome variable of the main data model is not included within the additional models (which are the imputation models). Further, once a variable is used as an outcome in any of the imputation models, such a variable is no longer included in the specification of any further additional imputation models. The order in which the imputation data models are built is not important, as the BUGS script is declarative and not procedural. In addition, as the variables are assumed to be statistically exchangeable, the order in which they are listed in unimportant.

D.3 Specification of priors

A likelihood function is specified for each of the outcome variables. Sets of parameter priors are also specified for the explanatory variables. Recall that the explanatory variables could be used as outcome variables in imputation models. The choice of particular prior is typically through a sensitivity analysis. During this sensitivity exercise, various potential prior distributions are explored to establish which one yields a better model by executing and by efficient convergence. Note that the prior's distributions also have parameters that define the distribution. On occasion, these parameters are also distributions in themselves, and a further sensitivity analysis is conducted to enable a choice of these.

D.4 Initial conditions of the models

This Appendix D.4 includes the list of initial conditions specified for the model system parameters, as well as the initial values for first elements of categorical variables. A range of initial values are typically tried out prior to establishing which ones facilitate efficient convergence of the particular Bayesian model.

D.5 R-script code that wraps the Bayesian models

The R-scripts code that wraps all the components of the Bayesian model is included here. The wrapper manages the data pre-processing, and also contains the priors, initial conditions specified for the system parameters, initial values for first elements of categorical variables, and variables within the main and missing data models.

D.6 Code for the 'dsum' construct implemented within JAGS

The 'dsum' construct is implemented within JAGS (Kruschke, 2014, Plummer, 2003) environment, and enables constraints to be applied to augmented flows to group stations on the railway network. The 'dsum' is a particular feature of the JAGS software for Bayesian modelling.

D.7 Variabls considered for the imputation models

Outcome explanatory (O) or Exposure predictor (E) variable	**Model 1: flows**	**Model 2: journey factor**	**Model 4: ticket validity**	**Model 5: group stn. ind.**
Passenger count (Freq)	Y			
Origin - station entry	Y	Y	Y	Y
Destination – station exit	Y	Y	Y	Y
Origin – house/ other Postcode	Y	Y	Y	Y
Destination – work/ other Postcode	Y	Y	Y	Y
Distance (journey, access, or egress)	Y	Y	Y	Y
Cost (monetary, or time)		Y	Y	
Purpose	Y	Y	Y	Y
Gender	Y	Y	Y	Y
Age	Y	Y	Y	Y
Income	Y	Y	Y	Y
Household Type/children/cars	Y	Y	Y	Y
Ethnicity	Y	Y	Y	Y
Access mode to railways		Y	Y	Y
Egress mode from railways		Y	Y	Y
Daily Rates of ticket use – journey factor	Y	Y		
Ticket validity period (days/months)	Y	Y	Y	
Group station flow/indicator	Y	Y	Y	Y
Number of stops	Y	Y	Y	Y
Travel time (gross/actual)	Y	Y	Y	Y
Stn. arrivela/actual departure - time	Y	Y	Y	Y
Station access/egress distance	Y	Y	Y	Y

Note: Model 3 has been excluded because it was not adopted to produce the results presented. Model 6 is excluded as it is simply the 'dsum' construct included in the CD accompanying the thesis.

Appendix E
CAR-BYM and Geospatial Kriging models

E.1 Summary of CAR-BYM rail-heading model

The CAR-BYM model created in this research and implemented in open BUGS software (Spiegelhalter et al., 2007, Sturtz et al., 2010) is included her to enable a reproduction of the results reported in this thesis. The specifications of the priors and initial values are also included. The JAGS implementation of the CAR-BYM model consists of two parts, the R-script wrapper which contains the data pre-processing stages as well as the initial conditions of the parameter values and unobserved data values.

E.1.1 Model of likelihood functions and priors

The specification of the likelihood functions, the covariates, and the prior specifications are included here.

E.1.2 R-script including initial conditions

The R-software script developed to implement the Bayesian CAR-BYM model is included here, and it included the specification of the initial conditions and information to run the model.

E.2 Summary of Geospatial Kriging model

The specification of the likelihood functions, the covariates, and the prior specifications for the geostatistical kriging model used to investigate the impact of a new station at Kirstall Forge are included here.

E.2.1 Model of likelihood functions and priors

The specification of the likelihood functions, the covariates, and the prior specifications are included here for the disaggregate Kriging model.

E.2.2 R-script including initial conditions

The R-software scripts developed to implement the Kriging (standard and hierarchical) models are included here, and it includes the specification of the initial conditions and attributes specified to run the model.