Estimating Movement from Mobile Telephony Data

John Doyle

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

A thesis submitted in partial fulfilment
of the requirements for
Doctor of Philosophy

Department of Electronic Engineering
National University of Ireland Maynooth
Ireland

Head of the Department: Dr. Ronan Farrell

Research Supervisors: Prof. Seán McLoone and Dr. Ronan Farrell

“Everything that happens in your life helps to make you what you are today.
Your past is your future.”

Jim McGuinness

Declaration Of Authorship

I hereby certify that this thesis, which I now submit for assessment on the programme of study leading to the award of PhD has not been submitted, in whole or part, to this or any other University for any degree and is, except where otherwise stated the original work of the author.

Signed: ....

Date: May 5, 2014

Abstract

Mobile enabled devices are ubiquitous in modern society. The information gathered by their normal service operations has become one of the primary data sources used in the understanding of human mobility, social connection and information transfer. This thesis investigates techniques that can extract useful information from anonymised call detail records (CDR). CDR consist of mobile subscriber data related to people in connection with the network operators, the nature of their communication activity (voice, SMS, data, etc.), duration of the activity and starting time of the activity and servicing cell identification numbers of both the sender and the receiver when available.

The main contributions of the research are a methodology for distance measurements which enables the identification of mobile subscriber travel paths and a methodology for population density estimation based on significant mobile subscriber regions of interest. In addition, insights are given into how a mobile network operator may use geographically located subscriber data to create new revenue streams and improved network performance. A range of novel algorithms and techniques underpin the development of these methodologies. These include, among others, techniques for CDR feature extraction, data visualisation and CDR data cleansing.

The primary data source used in this body of work was the CDR of Meteor, a mobile network operator in the Republic of Ireland. The Meteor network under investigation has just over 1 million customers, which represents approximately a quarter of the country's 4.6 million inhabitants, and operates using both 2G and 3G cellular telephony technologies.

Results show that the steady state vector analysis of modified Markov chain mobility models can return population density estimates comparable to population estimates obtained through a census. Evaluated using a test dataset, results of travel path identification showed that developed distance measurements achieved greater accuracy when classifying the routes CDR journey trajectories took compared to traditional trajectory distance measurements. Results from subscriber segmentation indicate that subscribers who have perceived similar relationships to geographical features can be grouped based on weighted steady state mobility vectors. Overall, this thesis proposes novel algorithms and techniques for the estimation of movement from mobile telephony data addressing practical issues related to sampling, privacy and spatial uncertainty.

Acknowledgements

First and foremost, I would like to extend my sincere gratitude and thanks to my research supervisors Dr. Ronan Farrell and Prof. Seán Mcloone for their continuous support throughout my PhD. Each in their own unique way, acted as both a mentor and a friend. Their encouragement, guidance and ability to impart knowledge of mathematical and engineering principles where vital in the completion of this body of work. Particular thanks are also due to Dr. Peter Hung whose mentoring ability was only out done by his editorial masterclasses, friendship and domain expertise. I thank all of you for your help, I could not have finished my thesis without you.

I would also like to extend my gratitude to Dr. Damian Kelly, Dr. Muhammad Tahir and Dr. Tim McCarthy for their contributions and involvement in this body of work.

The Research presented in this thesis was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan and by the Irish Research Council for Science, Engineering and Technology through an IRCSET postgraduate scholarship in partnership with ESRI Ireland. I would like to gratefully acknowledge this support, in particular Michael Byrne of ESRI Ireland.

I would also like to gratefully acknowledge the support of Meteor for providing the data used in this thesis, in particular John Bathe and Adrian Whitwham.

Ordinance Survey Ireland are also gratefully acknowledged for the provision of geographical region polygons and transportation network vectors.

A significant contributor to my continued enjoyment of the past few years has been the great community spirit that exists within the Electronic Engineering Department at the National University of Ireland Maynooth. Thank you to all of the staff, especially Joanne Bredin and Ann Dempsey. For their priceless IT and hardware support, I would also like to thank John Maloco, Denis Buckley and James Kinsella.

Thanks to the NUIM postgrads and espically my colleagues and friends in the Callan Institute for the banter, support and guidance, espically Dr. John Dooley, Dr. Grzegorz Szczepkowski, Dr. Alvaro Palomo Navarro, Dr. Tomasz Podsiadlik, Han Su, Emmett Carolan, Tony Keenan (extra thanks for the lifts to work, priceless), Keith Finnerty, Mary Larkin, Aidan McDermott, Diarmuid Collins, Prakash Srinivasan, Felix Wu and Sarah Adel. Also, a special thanks to Joanna O'Grady for all the help throughout my time within the Institute. I would also like to extend my gratitude to my colleagues in StratAG, in particular Dr. Jan Rigby and Melina Lawless for their support in completing this research.

To my friends outside college, thank you for the understanding, advice and sanity checks. In particular, a special thanks to Darren, Brendan, Gabhan, Seamus, Paul, Ciara, the Rage and Luo.

To Ursula, you are my best friend and my rock. Your love, encouragement and support are the main reasons I was able to finish this thesis. No matter how tough things got, you were always there. Thank you for being you, for your hugs, kisses, motivation, your understanding and for your constant loving support. I will never forget everything that you have done. You were always there when I needed someone to talk to or when I was having a bad day or week. Even though you were busy with your own PhD you would not hesitate in being there for me and for us. You mean everything to me, and I am forever yours. I would also like to extend a special thanks to Ursula's family for their support, encouragement and help over the last few years.

Finally, I would like to thank my own family for all their support over the last few years in particular. My parents Eddie and Kathleen for your unconditional love, your understanding and the constant words of encouragement. To my sisters Emma and Catherine, my extended family Brendan and James and my nieces and nephew Kathlyn, Jasmine, Amber Rachel and Seamus, thank you for the constant laughs and much needed distractions. For all the support you have given and for being there when needed, even if it meant moving tonnes of equipment half way around the country, driving hours to fix car related issues or the time spent sorting out issues, thank you all.

Dedicated to both Ursula and my parents with love.

Contents

Publications Arising From This Thesis

The following papers have been published based on the work presented in the thesis.

- **Doyle, J.**, Hung, P., Farrell, R. and McLoone, S. (2013). Population Mobility Dynamics Estimated From Mobile Telephony Data, under review, Journal of Urban Technology (JUT).

- **Doyle, J.**, McLoone, S., Hung, P. and R. Farrell (2012). Estimating Movement From Mobile Telephony Data, Mobile Tartu, 2012.

- **Doyle, J.**, Hung, P., Kelly, D., McLoone, S. and Farrell R. (2011). ”Utilising Mobile Phone Billing Records for Travel Mode Discovery”, Proc. 22th IET Irish Signals and Systems Conference, 2011.

- Kelly, D., **Doyle, J.**, and Farrell, R. (2011). ”Analysing Ireland’s Social and Transport Networks using Sparse Cellular Network Data”, Proc. 22th IET Irish Signals and Systems Conference, 2011.

- **Doyle J.**, McLoone S., McCarthy T. and Farrell R. (2010). ”Topography of Irish Mobile Telephony Activities: Visualising Human Dynamics on a Macro Scale”, GeoVA(t) - Geospatial Visual Analytics: Focus on Time Workshop, AGILE, 2010.

- **Doyle, J.**, Farrell R., McLoone, S., McCarthy T., Tahir M. and P. Hung (2009). ”Utilising Mobile Phone RSSI Metric for Human Activity Detection”, Proc. 20th IET Irish Signals and Systems Conference, 2009.

- **Doyle, J.**, Farrell R., McLoone, S., McCarthy T. and P. Hung (2009). ”Extracting Localised Mobile Activity Patterns from Cumulative Mobile Spectrum RSSI”, China-Ireland International Conference on Information and Communications Technologies, 2009, pp. 75-82.

List of Abbreviations

2G Second-Generation Wireless Telephone Technology.

3G Third-Generation Wireless Telephone Technology.

4G Forth-Generation Wireless Telephone Technology.

A-GPS Assisted Global Positioning System.

AOA Angle of Arrival.

APN Access Point Name.

BSC Base Station Controllers.

BTS Base Transceiver Station.

CDMA Code Division Multiple Access.

CDR Call detail Record.

CSO Central Statistics Office Ireland.

CSV Comma-separated values.

DTW Dynamic Time Wrapping.

ED Electoral Division.

EDGE Enhanced Data rates for GSM Evolution.

eUTRAN Evolved UMTS Terrestrial Radio Access Network.

eNode B Evolved Node B is the eUTRAN equivalent of the UTRAN Node B.

E-OTD Enhanced Observed Time Difference.

ETP Estimating Travel Paths.

FDM Frequency Division Multiplexed.

FDMA Frequency Division Multiple Access.

GERAN GRAN with the addition of EDGE packet radio services.

GPRS General Packet Radio Service.

GPS Global Positioning System.

GRAN GSM Radio Access Network.

GSM Global System for Mobile Communications.

GTP Generated Travel Path.

HLR Home Location Register.

IMSI International Mobile Subscriber Identity.

KDE Kernel Density Estimate.

LAI Location Area Identity.

LBS Location Based Services.

LCSS Longest Common Subsequence.

LOS Line of Sight.

LPCC Longest Common Subsequence modified using PCC.

LTE Long Term Evolution Wireless Telephone Technology.

LVCP Longest Common Subsequence modified using VCP.

MD Mobile Telephony Enabled Devices.

MME Mobility Management Entity.

MPCC Modified Probabilistic Cell Connectivity.

MSC Mobile Switching Centre.

MS Mobile Station.

MVCP Modified Virtual cell Path.

NLOS Non Line of Sight.

Node B UMTS equivalent of a Base Station Transceiver.

NUI National University of Ireland.

OCI Other cell Interference.

OFDMA Orthogonal Frequency Division Multiple Access.

OSI Ordnance Survey Ireland.

OTD Observed Time Difference.

O-TDOA Observed Time difference of Arrival.

PCC Probabilistic Cell Connectivity.

RNC Radio Network Controller.

RSSI Received Signal Strength Indication.

SFTP Secure File Transfer Protocol.

SGSN Serving GPRS Support Node.

SIM Subscriber Identity Module.

SMS Short Message Service.

SS Signal Strength.

STB Space Time Bead.

STP Space Time Prism.

TA Timing Advance.

TDMA Time Division Multiple Access.

TDOA Time Difference of Arrival.

RFID Radio-frequency Identification

UTMS Universal Mobile Telecommunications System.

UTRAN UMTS Terrestrial Radio Access Network.

VCP Virtual cell Path.

VLR Visitor Location Register.

WCDMA Wideband Code Division Multiple Access.

List of Symbols

α	A scalar to balance the learnt mobility patterns summarised by a mobility Markov chain with the influence of random transition probabilities
$\bar{H}$	Normalised H
$\bar{h}_{ij}$	Normalised h_{ij}
$\bar{N}_i$	The number subscribers living within an individual ED i
$\hat{D}$	Normalised from of D
Υ_u	The transition intensity matrix of the uth subscriber
A	A diagonal matrix used to convert H to D
a_i	Is the spatial area of ED_i
D	A matrix which relates ED building density to cell coverage regions
H	A matrix which relates the number of buildings in each ED to cell coverage regions
h_{ij}	The number of homes from ED i assigned to region of interest j
M	The number of EDs
N_s	The number of states in a transition probability matrix

N_u	The number of subscribers
N_j	The number of estimated subscribers living in a region of interest
P	A transition probability matrix
P_u	The transition probability matrix of the *u*th subscriber
p_{ij}	The conditional probability that a process will transition from state i to state j
Q	A modified a Markov chain

List of Figures

List of Tables

CHAPTER 1

Introduction

The ubiquitous nature of technology in modern human civilisation coupled with advancements in data acquisition techniques has resulted in the quantity of data related to human activity rapidly expanding. Typically, information which corresponds to human activity may be characterised as survey based, passive, activity based and device based. The availability of data, accuracy, scalability, cost and information content related to each approach varies dramatically.

Survey based approaches gather human activity data, preferences and behavioural data through the information entered by people partaking in a survey. Surveys such as national census [2], household surveys [3, 4] and customer satisfaction surveys [5] play an important roll in formation of government policy [6], regional planning [7] and corporate governance [8]. Studies on travel behaviour [9, 10], marketing [11, 12] and health [13, 14, 15] demonstrate how surveys may be used to gauge public opinion, behaviour and mobility.

Passive sensing approaches infer human activity through static sensors and the loads observed by service networks. Examples of passive sensors include flow counters [16, 17], traffic cameras [18, 19, 20] and sensing devices [21, 22, 23, 24]. The spatial and temporal resolution achieved by using these devices is often dictated by the spatial distribution of the sensors and their respective coverage areas. Mobile device activity [25], WiFi activity [26] and the flow of Bluetooth enabled devices [27] have also been used to measure both the spatial and temporal dynamics related to human activity. The tracking of bank notes [28], spatial variations in the amount of available public bikes [29], sensor GPS traces [30] and the tracking of travel cards [31] also demonstrated that passive sensing approaches could be used to observe human mobility behaviour.

Activity based approaches record when, where and how people interact with services and applications. Examples of activity based monitoring includes among others the making of phone calls [32, 33], tracking behaviour from retail membership reward cards [34, 35] and the tracking of consumer purchases [36, 37, 38]. The activity data sourced from services such as Twitter, Foursquare and Flicker have also be used to gauge public opinion [39, 40, 41], study human mobility [42, 43, 44, 45] and social connection [46].

The behaviour and location of people may also be recorded by sensing devices [47]. Sensing platforms [48, 49, 50, 51, 52], sensors placed on mobile cellular devices [53, 54], radio-frequency identification (RFID) tags [55, 56], Bluetooth monitors [57] and global positioning system (GPS) logging devices [58, 59, 60, 61] have been used in applications related to among others health, land use, mobility and intelligent transportation.

Each collection methodology has associated uncertainty with respect to the quality of gathered human activity information. For example, the accuracy of survey based approaches may be influenced by the memory of each individual partaking in the survey [62], sensor data is affected by the accuracy of each device [63] and activity based approaches have sampling related issues [64]. As a result, the selection of which collection method to use given accuracy requirements, scalability, associated cost and desired information needs careful consideration. Consideration should also be given to the privacy of monitored individuals, as studies which infringe on this right may face legal reprimand.

Typically, the cost associated with carrying out a survey can be prohibitively expensive. As a result, large surveys such as censuses tend to be carried out infrequently. However, the information which may be gathered can be extremely detailed as people can enter complex information about their preferences, behaviour and relationships. The information which may be extracted from a passive sensing application is dependant on the sensing device used. Large urban and national scale observations are limited by the cost of individual sensors and infrastructure requirements. As a result, passive sensing applications are better suited to monitoring human activity at local scales. Similarly, device based studies rely on the cooperation of each individual carrying the sensing device to share sensor information, thus limiting the scalability of such studies.

Activity based approaches may be suited to macro scale analysis if the service network or application from which the activity measurement was taken is ubiquitous in society. However, the information which may be gathered is limited by the type of service/application used. For example, WiFi usage patterns do not reveal the gender of the person using that service or their music preferences. Likewise, mobility studies which require both high spatial and temporal sampling might not be suited as temporal sampling resolution is dictated by the activity profile of each individual person and spatial accuracy is limited to the underlying network of the service/application.

1.1 Motivation

Measuring the movement of people is a fundamental activity in modern society. The ability to monitor movement insures that transportation services are able to function, planning authorities can design adequate infrastructure and governments can implement national policies. Personal location data is also the primary data source used in the delivery of mobile telecommunications [65] and location-based services (LBS) [66].

The research presented in this thesis uses human activity data to develop applications related to different aspects of human movement behaviour. The primary data sourced used is anonymised CDR from Meteor, a mobile network operator in the Republic of Ireland. Call detail records (CDR) consist of user information relating to people in connection with the network operators, the nature of the communication activity (voice, SMS, data, etc.), duration of the activity, starting time of the activity and servicing cell identification numbers of both the sender and the receiver when available. As a result, assuming users carry their mobile devices most of the time, it has the potential to be a low cost scalable source of human activity data.

While such activity data has a wide variety of applications, this research focuses on population density estimation, travel path identification and marketing insights through CDR metrics. This is motivated by the host of potential practical applications which include, among others, utility load forecasting and dynamic transportation services. The ability to deliver intelligent geographically located marketing applications to create new revenue streams for mobile operators is also an attractive proposition for in-dept investigations.

Likewise, dynamic population estimation can supplement tradition national census. Current population estimation research efforts using mobile generated data often require the estimation of mobile subscriber home locations which can be computationally intensive and may have privacy related issues [67, 68, 69, 70]. As a result, there is a need for different approaches which can provide population density measurements which are both computationally efficient and privacy preserving. Research presented within this thesis addresses such concerns by obtaining a direct measure of population density through the steady state vector of a modified Markov chain mobility model characterising the regional transitions of Meteor's customers.

The cost associated with transportation surveys motivates the requirement for low cost and scalable alternatives. Exploiting mobile network data is an attractive proposition in this regard [71]. However, due to the lack of trip metadata, spatial uncertainty and temporal sampling issues, it is difficult to relate CDR positional estimates to transportation related features. Addressing these issues, similarity metrics are developed to quantify the resemblance between CDR trajectories and known travel paths between regions of interest.

1.2 Privacy

Over the last decade, the boundaries and content of what is considered private have been the subject of much debate and legal challenge. As defined by Westin [72] "Privacy is the claim of individuals, groups or institutions to determine when, how, and to what extent information about them is communicated to others, and the right to control information about oneself even after divulgating it" [71]. Typically what is considered private differs between various groups and individuals, and is often challenged under the title of "public interest".

The ubiquitous nature of wireless technologies and their subsequent impact on privacy has directed legislators to devise various laws and regulations which govern how content sourced from such devices may be used and handled [73]. Within the European Union, there are several pieces of legislation which address privacy. Article 7 of the Charter of Fundamental Rights of the European Union (2000/C364/01) [74], states 'Everyone has the right to respect for his or her private and family life, home and communications'. Directive 95/46/EC [75] of the

European Parliament and of the Council outlines a framework for the protection of individuals with regard to the processing of personal data and on the free movement of such data. Directive 2002/58/EC [76] of the European Parliament and of the Council, concerns the processing of personal data and the protection of privacy in the electronic medium of communications. Directive 2006/24/EC [77] of the European Parliament and of the Council, instructs providers of electronic communications services and networks to keep traffic data record related to telephony communication and emails for a period of six months to two years, depending on the Member State. The traffic data includes the information which is required for identifying the originator and the recipient of phone calls (including Internet telephony), SMS and emails, together with information on the time, date, and duration of these communications [73].

As a result of the aforementioned directives and the commercial interests of service providers, access to such data is difficult to obtain. In the telecommunication sector, access to data governed by Directive 2006/24/EC has typically only been made available to a number of selected research institutions and commercial entities after contractual agreements are put in place which govern its use. Privacy issues stemming from such collaborations customarily arise when the tracking of people or goods transported are addressed [71]. To ensure that telephony data does not breach current regulations on data protection, such information should be received and handled in an aggregate and anonymous manner, which maintains user privacy.

Typically, user anonymity is addressed by a hashing of the user's unique MSISDN code. A MSISDN is a uniquely identifiable code which links to a person's subscription on a mobile cellular network. Such hashing guarantees that a user's identity is not directly observable. However, research has shown that through aggregation with external data sources and prior knowledge, a user may still be identified [78, 79, 80]. As a result, techniques have evolved which aim to hide user identity through forms of aggregation [81, 82, 83, 84, 85]. The research detailed within this thesis addresses privacy through a hashing of MSISDN codes and suitable aggregation. Note, no attempt to aggregate the data with external sources has been made or allowed. The details of additional steps taken are outlined in subsequent chapters.

1.3 Thesis Contributions

The general focus throughout this thesis is on the development of applications for large scale mobility estimation through the use of mobile telephony call detail records (CDR). Methodologies are developed which enable applications such as population estimation, travel route discovery and geographical marketing. In this context, the main contributions of the research presented in this thesis are as follows:

- The development of a novel methodology and distance measurements which enables the identification of mobile subscriber travel paths.
- The development of a novel methodology for population density estimation based on significant mobile subscriber regions of interest.
- Insights into how a mobile network operator may use subscriber generated data to help create new revenue streams and improved network performance.

Other minor contributions of this thesis include:

- A methodology for CDR feature extraction, data visualisation and cleansing techniques.
- A novel procedure for simulating journey trajectories along known travel paths.
- A novel procedure for constructing Generating Travel Paths (GTP) from CDR journey trajectories, where a GTP represents the likely path taken by a group of similar CDR journey trajectories which move between regions of interest, without prior knowledge of any underling travel routes.

1.4 Thesis Organisation

The remainder of the thesis is organised as follows:

Chapter 2 critiques the use of mobile cellular networks as a suitable sensing platform for monitoring human activity patterns. It starts by giving the reader a brief overview of modern mobile cellular networks and discusses the various techniques for mobile telephony data procurement, including an overview of current research using such data. It then details the development of a sensor used for the passive collection of client side mobile phone accumulative RSSI activity. The chapter concludes with a discussion of the evolution of mobile networks and how this might impact on human activity research directions into the future.

Chapter 3 describes the structure of call detail records (CDR) and the system used for data processing in this research. The development of cell coverage area models is also described, along with a suitable technique for mapping multiple cell coverage polygons to a single representative location covering a population centre. Then procedures for the extraction and visualisation of various CDR features are examined. The chapter concludes with a discussion on the taxonomy of possible applications which may be developed from the features extracted.

Chapter 4 details novel distance measurements (VCP and PCC) which enable the measurement of similarity between CDR journey trajectories and travel paths of interest. A comparison of established trajectory distance measurements and each of the proposed techniques is given, and it is demonstrated that both VCP and PCC achieve greater accuracy when classifying which route CDR journey trajectories took. Novel enhancements to the distance measurement Longest Common Subsequence (LCSS) are then given which improve the accuracy of LCSS with respect to CDR trajectory distance calculations. The CDR journey trajectories used in each comparative study, are generated using a novel procedure for simulated journey trajectories along known travel paths. Also detailed is a novel procedure used to construct Generated Travel Paths (GTP) of CDR journey trajectories. A GTP represents the likely path taken by a group of similar CDR journey trajectories which move between regions of interest, without prior knowledge of any underling travel routes. Finally, the chapter is concluded by discussing each of the topics covered.

Chapter 5 presents novel techniques for population estimation based on significant mobile subscriber regions of interest. The techniques use the steady state vector of a modified Markov chain mobility model which characterises the mobility of individual subscribers and national aggregated mobility, respectively, as a means of identifying the principle location of subscribers, thus providing a proxy for population density. Results show a high correlation between estimated population counts and a national census, which was carried out in 2011. A methodology for visualising the flow of people across the Republic of Ireland is also given, with insights into how the transition intensity observed may be used for event detection. The chapter is concluded by discussing the limitations and benefits of each technique presented.

Chapter 6 details initial work into the development of geographical marketing applications built on the outputs of previous chapters. This includes methodologies for identifying black spots related to high data rate subscribers, event mobility patterns and the segregation of subscribers based on their perceived links with geographical features of interest. The chapter is concluded by discussing how future research may build upon these initial findings towards the aim of fully commercialised applications.

Chapter 7 concludes the thesis with a summary of the work completed, contributions made to the field and the relevant areas of work which remain to be investigated.

CHAPTER 2

A Human Sensing Platform: Mobile Cellular Networks

In the last decade, mobile phones and mobile devices utilising mobile cellular network connections have become ubiquitous in modern society. In several developed world countries, the penetration of such devices has surpassed 100%. They facilitate communication and access to large quantities of data without the requirement of a fixed location or connection. As mobile phones and devices are mostly used by people, their activities and motion are indicative of the mobility pattern and cellular usage of the person using them. As such, the network of mobile phones and devices may be considered as a large scale distributed human activity sensing platform.

In this regard, exploiting the data collection capabilities of mobile devices is an attractive proposition. Over the last few years, various studies have demonstrated that information sourced from mobile device activities can be used to reveal space-time behaviour patterns relating to human mobility [86, 87, 88, 89], social structure [90] and land use [54, 67, 91, 92]. In all cases, it is noted that the type of analysis possible is strongly influenced by the underlying data collection methodology and the quantity of data available. Before actively acquiring such data, it is important to consider the structure and behaviours of a mobile phone network and how it may affect observations. Also, as mobile networks evolve to meet the future requirements of their customers, it is also important to consider how this will impact on future research directions.

This chapter begins by presenting a brief overview of modern mobile cellular networks. The procurement of mobile telephony data is then discussed in Section 2.2 and Section 2.3. Section 2.2 outlines the data sources available to mobile networks operators, and gives an overview of current research using such data. Section 2.3 follows with a discussion of alternative methods for obtaining mobile telephony data and details the development of a sensor used for the passive collection of client side mobile phone accumulative RSSI activity. Section 2.4 concludes the chapter with a discussion of topics covered and insights into the evolution of mobile networks and how this might impact on future research directions.

2.1 Mobile Telephony Networks

A mobile telephony network is a geographically distributed radio network that enables communication via voice, text or data between two or more devices [1, 93, 94, 95, 65]. At one particular time instance, each device has typically a wireless connection to one fixed-location transceiver, known as a tower. Each tower covers a service area, known as a cell, ranging from several square kilometres in rural areas to several hundred square metres in urban districts. Each device communication flow, including intra-cell communications, passes from the initiating device's connected transceiver through hierarchical network elements before being routed to the destination cell and subsequent receiving device, as depicted by Figure 2.1.

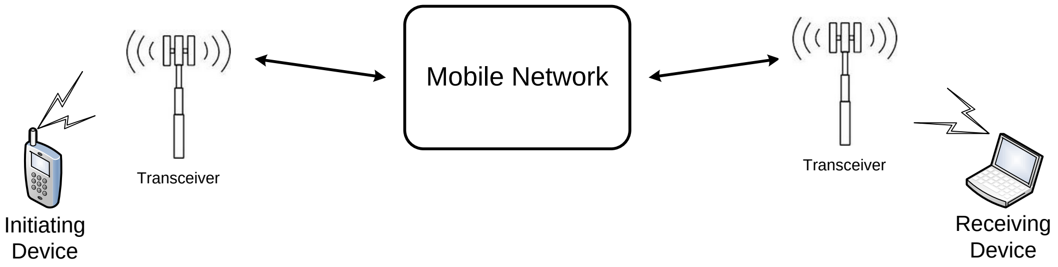


Figure 2.1: Simplified structure of a communication flow in a mobile telecommunication system.

A typical mobile network consists of a combination of second and third generation wireless telephone technologies (2G/3G), with newer systems employing long term evolution wireless telephone technology (4G LTE). A simplified hierarchical structure of the combined subsystems is depicted in Figure 2.2. For the purpose of human sensing, a mobile network may be divided into three main sections, namely the mobile subscriber layer, the radio access networks and the core network. The mobile subscriber layer is comprised of mobile telephony enabled devices or mobile stations (MS) which are subscribed to a mobile network. The radio access networks consist of radio transceivers used to transfer data from the MS to the core network. The core network is the central part of the mobile telecommunication network. It provides the services which enables mobility, communication and billing.

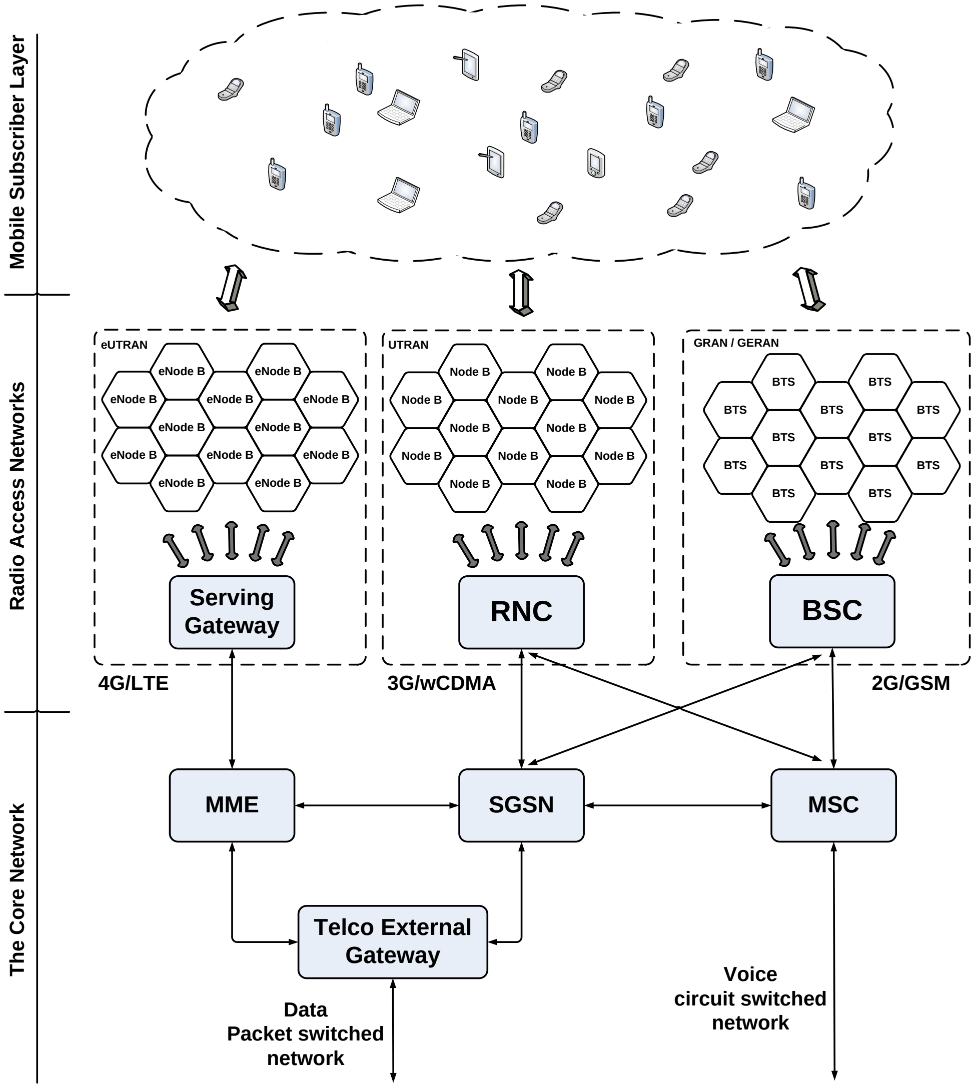


Figure 2.2: Simplified structure of a mobile telecommunication system.

Depending on the mobile communication standard employed, the radio access network types will vary between the 2G, 3G and 4G equivalents. A GSM radio access network (GRAN) consists of base transceiver stations (BTS) and base station controllers (BSC). A UMTS terrestrial radio access network (UTRAN) consists of Node B transceivers and radio network controllers (RNC). An evolved UMTS terrestrial radio access network (eUTRAN) is comprised of evolved Node B (eNode B) and serving gateways. The core network contains elements of the respective 2G, 3G, and 4G telephone technologies which includes among others, a mobile switching centre (MSC), serving GPRS support nodes (SGSN) and mobility management [65].

2.1.1 Multiple Access Techniques

Mobile telephony networks allow the simultaneous transmission and reception of communication between mobile devices, within a finite amount of radio spectrum. This is achieved by utilising several multiple access techniques, with the primary focus of permitting transmitting stations to communicate with receiving stations without any interference [65]. This increases overall network capacity as more communications can be facilitated within a limited amount of radio spectrum. The multiple access strategy employed varies between each generation of mobile telephony system, but each may be generalised by its primary approach, namely frequency, time or code division multiplexing. These strategies also effect network mobility management, a major component of cellular networks, as the locating strategies are often a function of the multiple access technique employed. The main approaches are briefed as follows:

- *Frequency Division Multiple Access (FDMA);* Here, individual channels or unique frequency bands are assigned to each mobile station on demand to users who require service. For the duration of the activity, no other device may use that channel.

- *Time Division Multiple Access (TDMA);* This approach divides the radio spectrum into time slots. In a similar fashion to FDMA, each time slot is assigned to an individual on an on demand basis, and is allocated to that user for the entire transmission.

- *Code Division Multiple Access (CDMA);* A spread spectrum technique, CDMA multiplies the narrowband message signal by a wideband signal called the spreading signal.

The spreading signal is a pseudo-noise code sequence that has a chip rate orders of magnitude greater than the data rate of the message signal [93]. Each active mobile device is assigned a spreading code, approximately orthogonal to all other codes, and may transmit simultaneously using the same carrier. For the receiver to be able to recover the original message it must know the spreading code applied. Decoding is achieved through a time correlation operation, where all other codewords appear as noise due to decorrelation [93].

- *Orthogonal Frequency Division Multiple Access (OFDMA);* This technique uses time sharing and dynamically assigned orthogonal subcarriers to provide multiple access to users. Users who require high data rates may be assigned a higher number of subcarriers compared to those who require low data rates.

For a comprehensive overview of these techniques see [1], [65] [93] and [94].

2.1.2 Spatial Coverage

As radio access network elements communicate wirelessly to devices present in the mobile subscriber layer, their transmissions suffer an effect known as path loss. Path loss refers to the amount of energy lost between transmission and reception of a signal. Assuming the use of an isotropic antenna for transmission, a propagated signal energy will expand over a spherical wavefront, so the energy received at an antenna a distance d away is inversely proportional to the sphere surface area, $4\pi d^2$ [94]. More precisely, the free space path formula, or Friss formula, is given as

$$P_r = P_t \frac{\lambda^2 G_t G_r}{(4\pi d)^2} \tag{2.1}$$

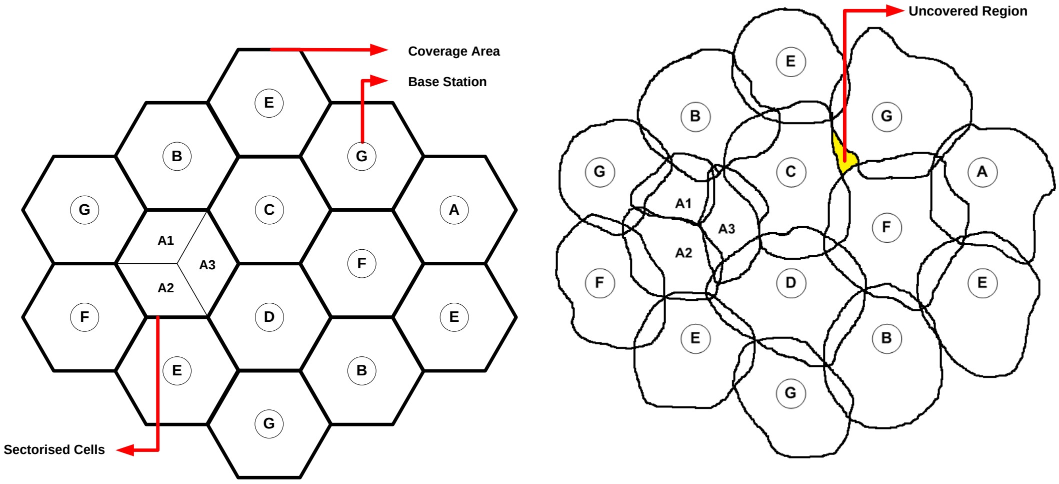
where P_r and P_t are the received and transmitted powers, λ is the wavelength and G_r, G_t refer to the receiver and transmitter gain, respectively.

Due to such effects, it is only possible to reliably communicate over some limited distance, given a maximum allowable transmit power. This allows transmitters to operate on the same frequency at the same time by virtue of being spatially isolated [94]. This effect is the theoretical basis for cellular mobile telephony systems as the overall capacity of a system increases as more simultaneous transmission are allowed to occur [94].

As such, the service area of a mobile telephony system is subdivided into smaller geographical regions. These smaller regions are commonly referred to as cells, and contain a single base station. To avoid interference between neighbouring cells, the transmit power level of each transceiver is regulated such that there is just enough to provide the required signal strength at the cell boundaries. Due to propagation path loss, the frequency channels they operate at may also be reused, as long as cells operating at the same frequency are spatially isolated. However, perfect spatial isolation cannot be achieved in practise, thus the rate of frequency reuse is determined such that the interference between cells is kept to an acceptable level [94]. This interference, which is known as other cell interference (OCI), still significantly impacts the performance of mobile systems. A common technique to reduce its effect is to sectorise cells, where sectorisation is achieved through the use of directional antennas [94]. As such, it is common for a cell to refer to an area covered by one sector, in which case a single base station site may have several associated cells [65].

Typical cell layouts are depicted in Figure 2.3. The hexagonal shape commonly associated with mobile telephony cells (Figure 2.3a), is an idealised depiction of coverage and does not accurately reflect actual cell boundaries. Instead Figure 2.3b more truly reflects their observed non-geometric shape with some areas not having the required signal strength for various reasons [65]. The spatial distribution of such cells is generally dictated by capacity requirements. In general, capacity can be increased by increasing the density of cells. This is achieved by turning down the transmit power to make cells smaller [94]. An operator may also use hierarchical cell structures (Figure 2.4), such as small cells, to increase network coverage or capacity in areas with very dense cellular usage [1].

Due to the fact that each mobile telephony standard is effectively frequency division multiplexed (FDM) in the radio spectrum, network planners design each network coverage layout independently. As a result BTS, Node-B and eNode-B may be mounted on a single tower, with each transceiver servicing the same particular geographical region in space. However, as mobile networks evolve and frequency reuse between standards becomes more prevalent, the layout of each network will be influenced, and sometimes limited, by the capacity requirements of other standards.



(a) Idealised

(b) Practical

Figure 2.3: Typical cell shapes. A-G refers to the frequency channel used by each individual base station.

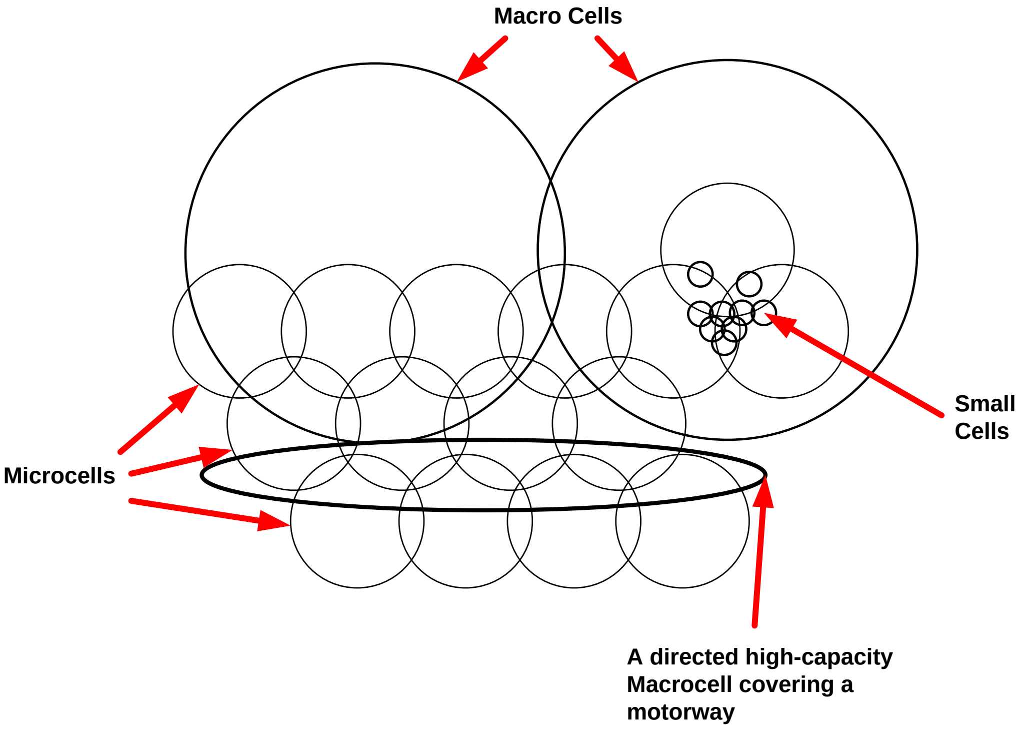


Figure 2.4: An example hierarchical cell structure [1]

2.1.3 Mobility Management

Mobility management is a major function of mobile networks. The aim of mobility management is to track where the MSs are, so that calls, SMS and other cellular services can be seamlessly delivered. In LTE systems, the MME is responsible for mobility management [94], which includes among others, functions such as tracking, handovers, paging and inter-cell interference coordination. There are two types of location registers used by GSM and UTMS networks [1, 93], namely the home location register (HLR) and the visitor location register (VLR). The HLR contains the permanent subscriber database which is registered to the operator's core network, while the VLR contains both registered subscriber information and details of devices which are roaming on the network. An entry is added to a HLR when a new mobile device or subscriber identity module (SIM) card is registered to the operator's network and remains static until subscription parameters are updated. Each HLR entry includes among others, the international mobile subscriber number (IMSI), MSISDNs, possible roaming restrictions, location area identity (LAI), MSC number and VLR number. The VLR contains similar information to the HLR, except information in the VLR is stored temporarily and contains the extra information on roaming customers.

To keep register information up to date, network operators require positional estimates of each mobile device. This has led researchers to investigate techniques which can be used to accurately locate mobile devices. The main approaches investigated are as follows:

- *Cell Identification;* This localisation technique uses the principle of proximity measurement, and involves identifying, communicating and locating the base station to which the mobile phone is connected. The located coordinates of the serving base station is then associated with the mobile device. The accuracy of such spatial information depends upon the physical topology of the mobile network, i.e. the size and coverage area of the cells.

- *Cell Identification + Timing Advance (TA);* GSM uses a combination of Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) of mobile stations for the efficient use of available spectrum [96]. Here each frequency slot is subdivided into eight time slots. Any one mobile device within a cell is assigned an individual frequency band and corresponding time slot in that band. For this protocol to work, packets sent by any mobile device must arrive at the base station in its assigned frequency band within its allocated time slot. The time the packet takes to travel from the mobile device to its serving base station will vary with the distance separating both. Base stations dynamically control when each device starts its transmission to try and insure that packets sent will arrive in their assigned window. This variable controlling of transmission start is known as Timing advance. Thus, the distance of a mobile device from a serving base station can be extracted from analysing the duration of the timing advance. However, this method is only used if the mobile user is 550 metres or more away from the serving base station. Adjustments are calculated depending on how many multiples of 500-550 metres the mobile user is from a base station.

- *Cell Identification + Signal Strength (SS);* in a similar fashion to the dynamic control of timing advance, the power at which a device may transmit is dynamically controlled by the serving base station. Since the attenuation in power experienced by a signal is a function of the distance travelled, base stations adopt a policy under which devices close are required to transmit with reduced power. This is to minimise the risk of devices close to base stations interfering with weak signals coming from devices further out. A base station implements this policy by monitoring the received signal strength indication (RSSI) of a mobile device. Once gathered, it then relays this information back to the device so that it can appropriately adjust its transmission power. Base stations try to maintain an optimal received signal strength to noise ratio for efficient communication. Thus, analysing the power or signal strength at which a transmitted signal is received allows inference of the distance between the mobile device and its serving base station.

- *Time Difference of Arrival (TDOA);* This is a triangulation technique, that can be performed by both mobile devices and mobile networks. Position is determined by triangulating the time needed for a packet to be sent from a device to three finely synchronised base stations and back. Problems exist as all transmitters and receivers in the system have to be precisely synchronised. A time stamp must also be inserted at the transmitting side in order for the measuring unit to discern the distance the signal has travelled.

- *Enhanced Observed Time Difference (E-OTD);* This is a TDOA-based location method based on the existing observed time difference (OTD) feature of GSM systems. OTD calculates the time difference between signals travelling from two different BTS to a mobile device. As mentioned above though, TDOA has synchronisation issues. Environmental conditions such as multipath fading and channel characteristics effect the perceived relative positions of such BTS and MS. As a result mobile network operators use location measurement units. Theses units compute the clock differences between base stations and then relay this information back into the network. BTS then transmit synchronisation information to various mobile devices [97]. Once synchronised, handsets equipped with software that locally computes location can calculate time differences and therefore distance from each base station, making triangulation possible.

- *Observed Time difference of arrival (O-TDOA);* This is a TDOA-based approach designed to operate over wideband-code division multiple access (WCDMA) networks. Effectively this is a WCDMA version of E-OTD [98].

- *Angle of Arrival (AOA);* This is a network localisation technique that uses a location scheme based on the principle of angulation [63]. The underlying principal of this method is a reversal of the concept of beamforming. The direction of a mobile device from a transceiver is determined by the wave incident upon an antenna array. Each antenna in the array makes a unique observation relating to the wave incident upon it, which generally relates to the difference in received phase of that wave. These differences in phase enable AOA to be calculated. A device is located by taking the intersection of vectors projected at angles determined by AOA from two or more transceivers.

- *Assisted global positioning system (A-GPS);* Here devices use both GPS and terrestrial cellular network localisation to obtain a geographic position [98].

The implementation of any one of the above techniques depends on the limitations imposed by the underlying mobile network structure. A more complete summary of techniques for both indoor and outdoor mobile device localisation is given by Sun *et al.* [97], Liu *et al.* [99], Kaemarungsi *et al.* [100], Pahlavan *et al.* [101], Hightower *et al.* [102], Jami *et al.* [103] and Sayed *et al.* [104]. Many of the aforementioned localisation techniques required line-of-sight (LOS) for accurate positioning, and may not be suited to localisation within an urban environment, where multipath non-line-of-sight (NLOS) communication is prevalent. For these reasons, advanced localisation techniques have been designed which take account for the existence of mixed LOS/NLOS conditions [105]. Such techniques have applied data fusion techniques to merge data from various sources [104], [106], exploited redundant measurements [107], combined analytical models with maps of measurements [108], [109], and used Bayesian methods to estimate a device's whole trajectory instead of estimating one position at a time [110], [111], [112], [113].

2.2 Mobile Operator Acquired Data

Modern mobile telephony networks routinely collect a wealth of information related to customer interactions in the context of their normal service operations. Functions such as connecting calls, delivering text messages via SMS or providing Internet access generate a huge amount of data which mobile network operators use for customer billing and service delivery.

Operator-based data sources include network bandwidth usage measurement logs which are typically measured in Erlang (in units of person phone use per-hour), handover records, locating area logs and call detail records (CDR). Handover records are recordings of migrations of a user from one servicing cell to another while in the process of an active call. Location updates area logs consist of periodic location updates relating to the set of cell towers which are prepared to service a particular mobile device at any given time. Call detail records (CDR) contain information about all interactions between a mobile phone network and their customers that are required for billing purposes. These contain anonymised user information relating to people in connection with the network operators, the nature of the communication activity (voice, SMS, data, etc.), duration of the activity, starting time of the activity and servicing cell identification numbers of both the sender and the receiver when available.

In typical telecommunication networks, such features are collected at the core network. When a subscriber activity occurs (i.e. customer makes a call, receives a SMS, etc.), their mobile device interfaces with either the BTS or Node B, the choice of which depends on current cell capacity utilisation, the subscriber's required data load, and their current 2G/3G connectivity. Note, at this stage high resolution user positional estimates may be collected via location triangulation or angulation, as discussed in Section 2.1.3, but the required information is not routinely stored and thus is typically not available. Instead, several of the activity logs including handover data logs, call detail records, network bandwidth usage measurement and user data quantities are stored at the relevant MSC or SGSN.

Whenever a change of user location area is detected, the MSC will initiate a transition update in either the location register HLR or VLR. This transition update may potentially provide more location-based data reflecting the mobility pattern of users compared to activity-based data. This is because of the generally higher sampling frequency by the nature of its information update. Also from the point of view of human mobility sensing, it does not suffer from the uncertainty associated with activity based updates. For example, in situations where users invoke activities only at starting and ending locations over a long distance journey, it may not be possible to estimate journey trajectories due to the lack of location-based data. Unfortunately, it is usually very difficult to obtain HLR and VLR data from mobile operators due to the lack of incentive for long-term storage. In contrast, mobile operators tend to treat activity-based call detail records with greater importance as it is required for legal compliance [73] and billing purposes. This explains the greater availability of CDR data for human mobility sensing.

Initiatives such as Data for Development [114], the Mobile Data Challenge [115] and CRAWDAD [116] have helped such datasets become more widely available in recent years. This has meant that there has been steady growth in the number of research groups which have gained access to human movement and behavioural data at urban and national scales. Ratti *et al.* [117, 118], Calabrese *et al.* [25, 119] and Horanont [120] each focused on the mapping of human activity. Ahas *et al.* [121] demonstrated that suburban commuter movements, tourist movement dynamics [32, 122, 123] and methods for home and work location estimation and population dynamics [68] could also be extracted from mobile telephony data sources. Tourist movements have also been studied by Kuusik *et al.* [124, 125], while alternative methods for home and work location estimation and population movement dynamics have been examined by Silm *et al.* [126], Calabrese *et al.* [127], Isaacman *et al.* [128], Kelly *et al.* [129] and Ranjan *et al.* [64].

Areas associated with mass urban activity may also be readily sourced from cell activity counts, as demonstrated by Reades *et al.* [67, 91], Andrienko *et al.* [130, 131], Becker *et al.* [132], Isaacman *et al.* [133], Vieira *et al.* [134] and Caceres *et al.* [92]. This type of work has generally focused on clustering areas of similar activity profiles. Clustering has also been applied to user groups with a focus on group movement patterns [135] and marketing [136].

Movement and mobility insights through mobile telephony data has also been a topic of discussion, most notably the works of Gonzàlez *et al.* [86] and Song *et al.* [87, 137] have provided insights on the basic laws governing human motion and limiting thresholds on human movement predictability. The range of human motion was quantified in Gonzàlez *et al.* using the radius of gyration [86]. This measures the overall range of an individual trajectory, and demonstrated a stark contrast between actual human motion and classical random walk models [138]. By measuring the entropy of individual trajectories, Song *et al.* [87, 137] showed that there was a potential predictability of 93% in user mobility across a mobile network operator's subscriber base, despite the significant differences in the travel patterns.

Mobility and movement prediction has also been the topic of works by Eagle *et al.* [69, 70], Park *et al.* [138], Couronne *et al.* [139], Kang *et al.* [140], Isaacman *et al.* [141], Vieira *et al.* [142], Lu *et al.* [143] and Phithakkinukoon *et al.* [144, 145]. Eagle *et al.* [69, 70] demonstrated the application and design of community structure algorithms that are appropriate for the identification of location clusters relevant to a mobile user's life. Validation of techniques was supplemented by Bluetooth beacons located in user homes. Mobility modelling algorithms were also developed using discrete Markov chains, for example by Park *et al.* [138], in which it was demonstrated that the approximation of user mobility through Markov chains reproduces the slow, sub-polynomial growth predicted by the evolution of the radii of gyration. Park *et al.* also discussed how the eigenvalues and eigenvectors of a Markov chain were related to an individual's mobility.

The availability of large quantities of human movement data has also been of interest in the transportation sciences. Various researchers have shown that mobile telephony networks can provide information which may convey transportation survey related parameters, including origin destination mobility, traffic speed, transportation mode, traffic volumes and home and work locations [71, 146, 147, 148]. Cell tower activity logs, in particular Erlang recordings, may be related to traffic parameters such as traffic density as it is linked to person occupation [149]. However, such logs are unsuitable for most other transportation survey parameters. Caceres *et al.* [150] demonstrated how handover and CDR could be used in the estimation of traffic volume. Bar-Gera [151] demonstrated that handover could be used in the estimation of traffic speeds and travel times. Caceres *et al.* [152] and White *et al.* [153] developed origin destination parameters utilising location updates and CDR, respectively, while Wang *et al.* [154, 155] and Doyle *et al.* [148] have both examined transportation mode inference techniques.

Other research efforts have examined the differences between Rural and Urban Societies [156] and produced agent-based models of epidemic spread [157], while Onnela *et al.* [90, 158], Kamola *et al.* [159] and Nanavati *et al.* [160] investigated the social network graphs produced from mobile subscriber interactions. The observed flow of network communication can also readily extracted from mobile telephony data sources. As such, it has been the subject of research by Lambiotte *et al.* [161], Krings *et al.* [162], Ratti *et al.* [163], Kelly *et al.* [129] and Walsh *et al.* [164].

2.3 Non-Operator Acquired Data

The natural source of mobile telephony data is from the data centres of mobile network operators. However, there is a number of difficulties when acquiring information in this manner, most notably the legal and privacy issues that prevent operators delivering such information to outside researchers. In addition, even with best efforts, there is no guarantee that data from theses sources is always available, complete or accurate. Network operators continually optimise their network throughout the day, using temporary towers. This adds a level of uncertainty to these fixed point measurements as network topologies become more dynamic. A more fundamental issue arises regarding spatial accuracy as the spatial resolution of the usage statistics is dependent on both the operator's network topology and base station hardware. As a result, approaches have emerged which aim to address these issues by placing either embedded software applications on the mobile devices to log data [53], or by constructing custom sensing platforms which monitor mobile devices in their vicinity [21, 22].

Using an embedded sensing application, Ahas and Mark [165] tracked the mobile phones of 300 users through a social positioning application. They combined spatio-temporal data from phones with demographic and attitudinal data from surveys to view a map of social spaces in

Estonia. MITs Reality Mining project [54, 166] also illustrated that it was possible to extract common mobility patterns from the activities of mobile phone users. The subjects were issued with mobile phones pre-installed with several pieces of software that recorded and sent research data on call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status. Other research efforts have examined social connection [167, 168], mobility [30, 169] and subscriber behaviour [170]. Another application to make use of embedded sensing applications is Google Maps® real time traffic estimator. The traffic information used in this feature comes from a combination of third party sources and data provided by Android users which have chosen to share information by opting into the “My Location” feature on Google Maps®.

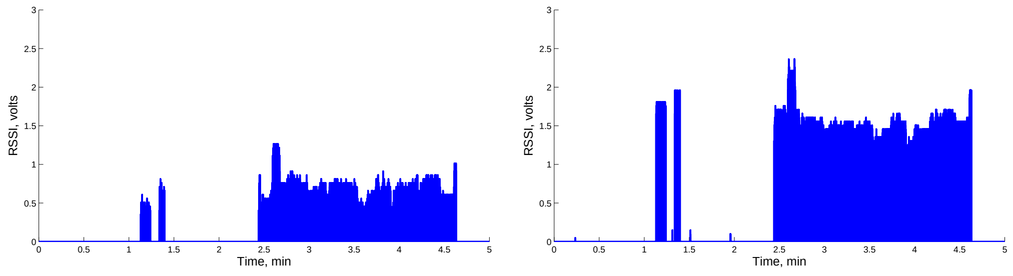
However, there are issues with embedded sensing applications. Aside from potential ethical concerns, embedded sensing applications require the cooperation of the device user to install intrusive software applications onto their mobile devices that enable the logging of information. This can have a limiting effect on the number of devices which can be sensed due to lack of cooperation from users. A compatibility issue may also arise as software applications on mobile devices are often platform dependent. As a result, researchers has started to develop custom-built sensing platforms for the passive collection of mobile telephony data.

Normal mobile network operational functions, such as device paging, are initiated when mobile devices are connected to a network. All mobile device activities occur in designated frequency bands and may be passively sensed using mobile receiver technology. Examples of custom-built sensing platforms include [21, 22] and [171]. Path Intelligence [171] have devised sophisticated sensing devices to return mobile device mobility patterns for the purpose of foot fall analysis in shopping centres. Their applications gather information from scanning the mobile phone frequency ranges and localising devices based on the characteristics of the radio signals observed. Typically, information which may be collected through custom sensing platforms includes in-range device positional estimates, in-range device count estimates and mobile spectral energy.

Doyle *et al.* [21, 22], highlighted the capabilities of cumulative received signal strength indications (RSSI) for the measurement of overall mobile device transmissions within the proximity of custom built sensing devices. The research, carried out on the north campus of the National University of Ireland Maynooth (NUIM), evaluated these sensing devices by undertaking two experiments. The first experiment investigated whether such sensors could detect spectral emissions from an SMS and phone call under controlled conditions. Results from this initial experiment are depicted in Figure 2.5 and Figure 2.6 respectively. The second experiment investigated each sensing device's capability to record mobile spectrum RSSI in an uncontrolled environment. Here, normal mobile device activity was observed from mostly the student population of NUIM over two consecutive time periods. Results from this experiment are given in Figure 2.7 and Figure 2.8.

The results of the first experiment indicate that the sensing nodes were capable of detecting normal mobile phone activity as the spectral energy associated with a text message and phone call were clearly visible. The second experiment performed on a non-controlled environment highlight events occurring close to each hour mark. These events relate to times where classes finished and started, and are as expected. These preliminary findings suggest that monitoring cumulative receiver signal strength measurements of mobile phone signals can be a valuable tool in gathering information for mobile phone usage independent of mobile phone operators on localised scales. However, these results are preliminary and the parameters of the temporal processing technique used, detailed in [21, 22], require further tuning.

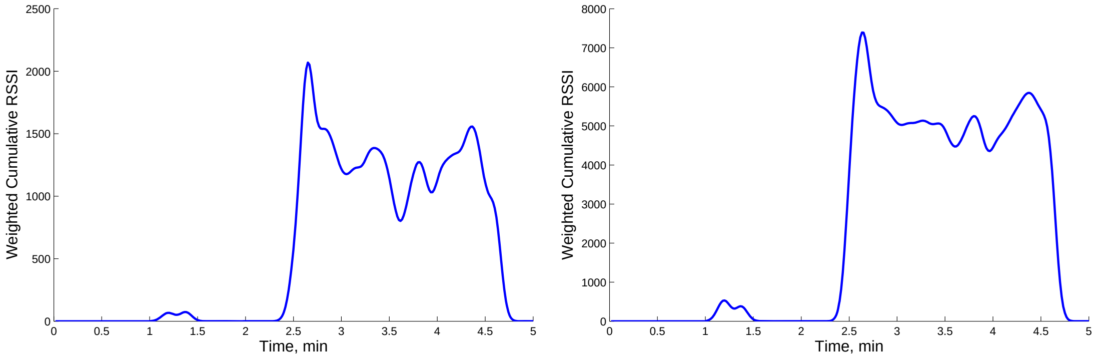
In the experimental setup, only an aggregated measure of spectral energy was recorded. Using more sophisticated sensors, the spectral energy of each unique device may be monitored. With this information, it is possible to obtain a more meaningful result for user occupancy. Such data could also be used to complement traditional techniques for mapping mobile device activity. For instance, one could use the network operator data, if available, to model the dynamics of a city or town, while localised RSSI data mapping could be employed to observe the dynamics of specific buildings or localised areas.



(a) Sensor A

(b) Sensor B

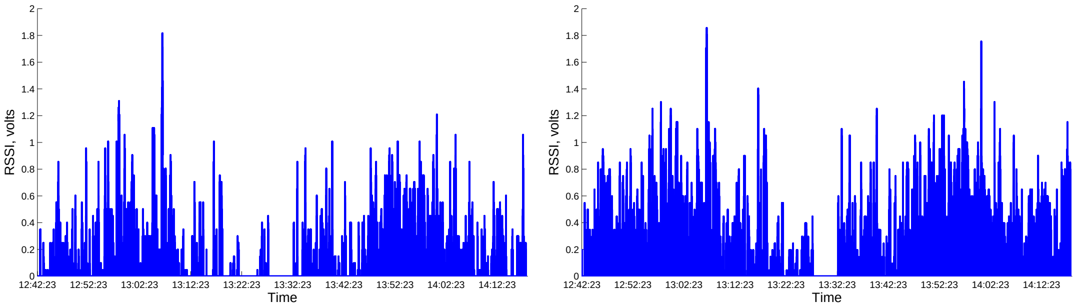
Figure 2.5: Observed mobile transmissions through recorded RSSI of mobile spectral energy from detecting sensors.



(a) Sensor A

(b) Sensor B

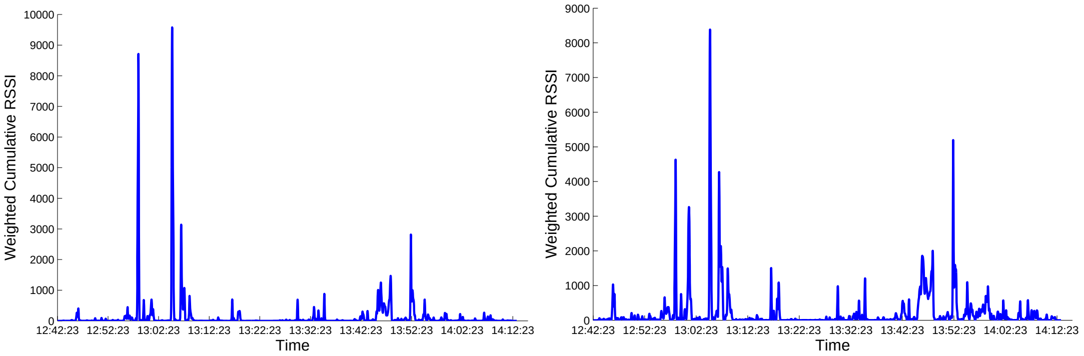
Figure 2.6: Weighted RSSI, highlighting time periods of high activity.



(a) Sensor A

(b) Sensor B

Figure 2.7: Normal mobile device activity from the student population of NUI Maynooth over consecutive time periods (1.5 hours).



(a) Sensor A

(b) Sensor B

Figure 2.8: Weighted RSSI, highlighting time periods of high activity.

2.4 Discussion

This chapter presented a brief overview of modern cellular telephony networks and discusses the various options available when gathering data from them. Supplementing this is a review of current research which uses mobile telephony data as a primary source of information. The development of a novel sensor which can be used for the passive collection of client side mobile phone accumulative RSSI activity is also briefly discussed.

Although the results from the initial investigation into RSSI collection have proved encouraging, the cost of building a distributed sensor network which could accurately monitor the client side mobile phone accumulative RSSI activity, even over a small geographical area, has proven to be prohibitively expensive. Alternatively, the costs associated with gathering data from a mobile network operator is minimal, as the infrastructure required to gather many of the logs outlined in Section 2.2 is in situ. While location accuracy is influenced by a mobile operator's cell tower topology, there exist opportunities to monitor mobile device activity at urban and national scales, which is not practical using only RSSI data collection.

As mobile phone networks evolve and Long Term Evolution (4G LTE) networks become more prevalent, the issues surrounding the location accuracy from mobile operator data will tend to diminish. This is due to the foreseen increased demand for mobile wireless broadband. As demand increases, the speed at which data is transferred will be required to increase substantially. Due to the effects of free space path loss, as outlined in Section 2.1.2, the most effective way of achieving this is to decrease area of cell coverage. This will result in an increased deployment of small cells throughout urban areas and towns.

As a result, the ongoing research presented in the remaining chapters of this thesis will concentrate on the use of operator provided mobile telephony data. The features that can be extracted from the call detail records (CDR) of Irish mobile phone provider Meteor are described in Chapter 3. The insights gathered from these features are then built upon in later chapters and form the basis of the primary contributions of this thesis.

CHAPTER 3

CDR Feature Extraction

There is a myriad of information which may be derived from mobile network operator sourced data that relates directly or indirectly to human activity. This chapter outlines the methodology used to extract and visualise a sample of such features from call detail records (CDR), cell tower information and subscriber registration data from one of the Republic of Ireland's cellular phone networks, Meteor. Insights into required procedures for data cleansing, cell coverage area modelling and cell clustering are also presented. Contributions include the estimation of the achievable distance a mobile device may travel over time, a novel activity spatial weighting function and a detailed discussion on the time variability of CDR trajectory sampling.

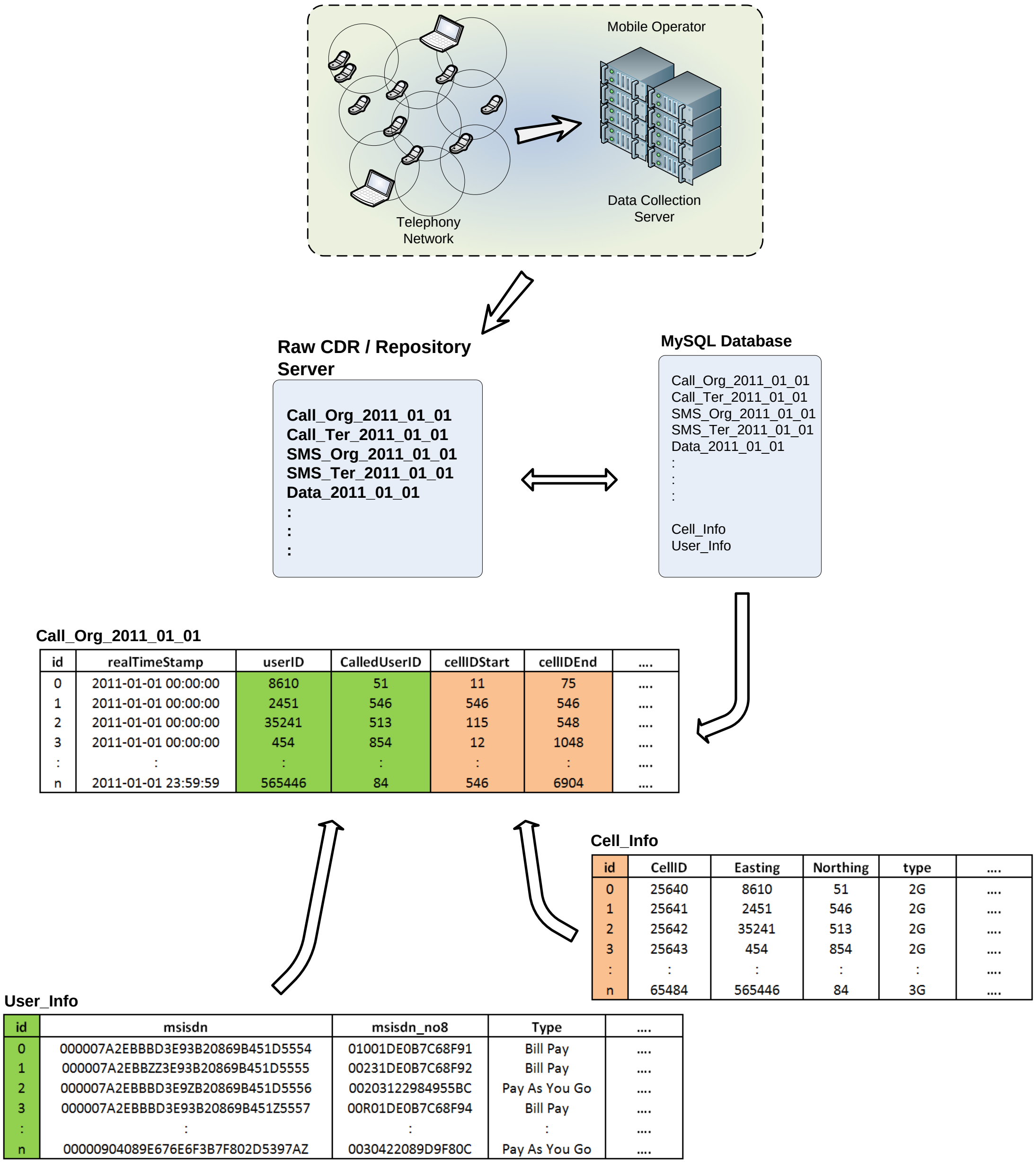
The Meteor network under investigation has just over 1 million customers, which represents approximately a quarter of the country's 4.6 million inhabitants, and operates using both 2G and 3G telephone technologies. The CDR are collected at the operator's MSC and SGSN and contain records related to voice calls, short message service (SMS) and data transfer. The available dataset consists of approximately three months of voice and SMS records from 09/11/2010 to 27/02/2011 and approximately two weeks of data records from 08/02/2011 to 27/02/2011.

The cell tower information provided contains geo-spatial coordinates in the Irish Grid Coordinate Reference System [172]. This coordinate system is used throughout this thesis, and uses the projections of Easting and Northing, which are in metre units from an origin located at a latitude of 53°30'00 N and a longitude of 8°00'00 W. Other cell information includes network type, transmitter azimuth, and the cell's associated MSC or RNC. The subscription information provided contains individuals anonymised MSISDN, subscription type (bill or prepay), year of birth and town of residency. Also included is information related to the number of upgrades and details on whether or not they have churned in from another network.

The voice calls and SMS records are split into originating and terminating files, while data logs contain information on mobile Internet sessions. The voice originating and terminating logs contain information on the time of each call, both caller and called subscriber's anonymised MSISDN, the duration of each call and the servicing cell towers of both caller and called subscribers at the start and end of each call when available. Similar information related to SMS activity is contained in the SMS originating and terminating logs. For each Internet session recorded in the data logs, information on the anonymised MSISDN, access point name (APN), session start time, duration of the session, servicing cell at the start of the session, quantities of data uploaded and downloaded, and servicing SGSN is collected. Note, cell information is only available for Meteor subscribed mobile stations.

The system architecture used to process the CDR consists of a repository server and three SFTP servers. The raw data in the form of CSV files were transferred from Meteor servers to the repository server. The repository server is used to hold all unprocessed data. Then SFTP servers are used to analyse the data. The data is transferred, preprocessed and stored in MySQL databases on these servers, with each table optimised for parameter extraction. From here an analyst may directly log on to the processing servers or remotely log in to the MySQL databases. An overview of this system architecture is illustrated by Figure 3.1. The data structures of call and SMS originating tables, call and SMS terminating tables and data session tables are given in Table 3.1 to Table 3.5 respectively.

The rest of this chapter is organised as follows. Section 3.1 describes the process used to model cell coverage areas. Section 3.2 details the identification of cell towers whose location information is incorrect. Section 3.3 then outlines procedures for the extraction and visualisation of various network activity metrics, while Section 3.4 details the methodology for the extraction of mobile subscriber trajectories. Such trajectories can be used to convey the aggregated flow of people between regions or towns. The grouping of cell towers into geographical coverage regions which service population centres is discussed in Section 3.5. Section 3.6 discusses how to summarise and visualise movements between such regions. The flow of information from one area to another is then examined in Section 3.7 and Section 3.8 describes how to extract social graphs from CDR. Finally, Section 3.9 concludes this chapter with a discussion on the taxonomy of possible applications which may be developed from the aforementioned features.



Call_Org_2011_01_01

id	realTimeStamp	userID	CalledUserID	cellIDStart	cellIDEnd	....
0	2011-01-01 00:00:00	8610	51	11	75	....
1	2011-01-01 00:00:00	2451	546	546	546	....
2	2011-01-01 00:00:00	35241	513	115	548	....
3	2011-01-01 00:00:00	454	854	12	1048	....
:	:	:	:	:	:	....
n	2011-01-01 23:59:59	565446	84	546	6904	....

Cell_Info

id	CellID	Easting	Northing	type	....
0	25640	8610	51	2G	....
1	25641	2451	546	2G	....
2	25642	35241	513	2G	....
3	25643	454	854	2G	....
:	:	:	:	:	....
n	65484	565446	84	3G	....

User_Info

id	msisdn	msisdn_no8	Type	....
0	000007A2EBBBD3E93B20869B451D5554	01001DE0B7C68F91	Bill Pay	....
1	000007A2EBBZZ3E93B20869B451D5555	00231DE0B7C68F92	Bill Pay	....
2	000007A2EBBBD3E9ZB20869B451D5556	00203122984955BC	Pay As You Go	....
3	000007A2EBBBD3E93B20869B451Z5557	00R01DE0B7C68F94	Bill Pay	....
:	:	:	:	....
n	00000904089E676E6F3B7F802D5397AZ	0030422089D9F80C	Pay As You Go	....

Figure 3.1: The system architecture used to process the CDR.

Table 3.1: CDR call originating table structure

Field	**Description**
id	Unique table row index
realTimeStamp	Formatted start time of the call
userID	Index link to the registration information for the subscriber making the call
CalledUserID	Index link to the registration information of the subscriber receiving the call
cellIDStart	Index link to the cell tower information of the cell servicing the caller when the call was initiated
cellIDEnd	Index link to the cell tower information of the cell servicing the caller when the call was terminated
TAC	The Type Allocation Code (TAC) of the mobile device making the call
callerMsisdn	The caller anonymised MSISDN
calledMsisdn	The called subscriber's anonymised MSIS
callTime	Un-formatted start time of the call
duration	The duration of the call
startCell	Cell tower ID of the cell tower which serviced the subscriber who made the call when the call was initiated
endCell	Cell tower ID of the cell tower which serviced the subscriber who made the call when the call was terminated

Table 3.2: CDR SMS originating table structure

Field	**Description**
id	Unique table row index
realTimeStamp	The formatted time at which the SMS was sent
userID	Index link to the registration information for the subscriber sending the SMS
CallerUserID	Index link to the registration information of the subscriber receiving the SMS
cellIDStart	Index link to the cell tower information of the cell servicing the subscriber who sent the SMS
TAC	The Type Allocation Code (TAC) of the mobile device sending the SMS
callerMsisdn	The anonymised MSISDN of the subscriber sending the SMS
calledMsisdn	The subscriber's anonymised MSISDN who is receiving the SMS
callTime	Un-formatted time when the SMS was sent
startCell	Cell tower ID of the cell tower which serviced the subscriber who sent the SMS

Table 3.3: CDR call terminating table structure

Field	Description
id	Unique table row index
realTimeStamp	Formatted start time of the call
userID	Index link to the registration information for the subscriber receiving the call
CalledUserID	Index link to the registration information of the subscriber making the call
cellIDStart	Index link to the cell tower information of the cell servicing the subscriber receiving the call when the call was initiated
cellIDEnd	Index link to the cell tower information of the cell servicing the subscriber receiving the call when the call was terminated
TAC	The Type Allocation Code (TAC) of the mobile device making the call
callerMsisdn	The caller anonymised MSISDN
calledMsisdn	The called subscriber's anonymised MSIS
callTime	Un-formatted start time of the call
duration	The duration of the call
startCell	Cell tower ID of the cell tower which serviced the subscriber who received the call when the call was initiated
endCell	Cell tower ID of the cell tower which serviced the subscriber who received the call when the call was terminated

Table 3.4: CDR SMS terminating table structure

Field	Description
id	Unique table row index
realTimeStamp	The formatted time at which the SMS was received
userID	Index link to the registration information for the subscriber receiving the SMS
CallerUserID	Index link to the registration information of the subscriber who sent the SMS
cellIDStart	Index link to the cell tower information of the cell servicing the subscriber who received the SMS
TAC	The Type Allocation Code (TAC) of the mobile device receiving the SMS
callerMsisdn	The anonymised MSISDN of the subscriber sending the SMS
calledMsisdn	The subscriber's anonymised MSISDN who is receiving the SMS
callTime	Un-formatted time when the SMS was received
startCell	Cell tower ID of the cell tower which serviced the subscriber who received the SMS

Table 3.5: CDR data session table structure

Field	Description
id	Unique table row index
realTimeStamp	The formatted time at which the data session started
userID	Index link to the registration information for the subscriber who is active
cellIDStart	Index link to the cell tower information of the cell servicing the subscriber when the session started
msisdn	The anonymised MSISDN of the subscriber who is active
datetime	Un-formatted start time of the session
apn	Access Point Name (APN) used by the mobile device
systemType	The system (2G/3G) the device is connected to
nodeid	SGSN id used in the session
accessPointNameNIapn	The Access Point Name (APN) used to identify an IP Packet Data Network (PDN), that the mobile data user communicates with
pdptype	The Packet Data Protocol used to transfer data, entry is empty for all CDR
uplinkBytes	Quantity of bits uploaded
downlinkBytes	Quantity of bits downloaded
duration	The duration of the session
TAC	The Type Allocation Code (TAC) of the mobile device active during the session
cellid	Cell tower ID of the cell tower which serviced the start of the session

3.1 Cell Coverage Regions

As outlined in Section 2.1.2 a BST, Node-B and eNode-B may be mounted on a signal tower, with each servicing various spatially overlapping geographical regions. Using the collective cell tower data, namely the geo-spatial coordinates and network type of each cell, it is possible to approximate idealised cell site coverage areas via Voronoi tessellation [173] for each mobile network of interest, where each centre represents a cell site location. Figure 3.2 depicts cell site Voronoi tessellations areas for 2G and 3G cell sites in our mobile network of interest. Note, the accuracy of the tessellation in approximating cell coverage areas is affected by channel characteristics, topography of the area and physical layer parameters which include transmitter frequency, tilt, height, and transmission power [65]. These factors have not been incorporated into this analysis, as the collection of such information is prohibitively expensive. As a result, it should be noted that the estimation technique applied does introduce some approximation error at a local level.

Figure 3.2 was produced using MATLAB® plotting functions. The Voronoi tessellation was produced using the MATLAB® function VORONOI, for which cell site locations of each network (2G/3G) were used as inputs separately. The function returns a polygon for each unique site location, thus cells with matching site location on the same network share the same site polygon. The county geographical regions polygons presented are sourced from Ordinance Survey Ireland (OSI) [174].

The coverage regions in Figure 3.2 are a reasonable approximation for cell site locations that lie within central locations, however, the absence of a limiting threshold for the size of coverage regions means that cells along coastal regions are poorly approximated. To address this a maximum cell site radius of 20 km and 15 km is introduced for 2G and 3G networks respectively. The choice of each limit reflects the realistic limit for communication with each standard given our network topology. Each site radius S_r is calculated as

$$S_r = \min \left\{ \sqrt{\frac{S_a}{\pi}},\ S_{max} \right\} \tag{3.1}$$

where S_a denotes the cell site coverage area and is given by

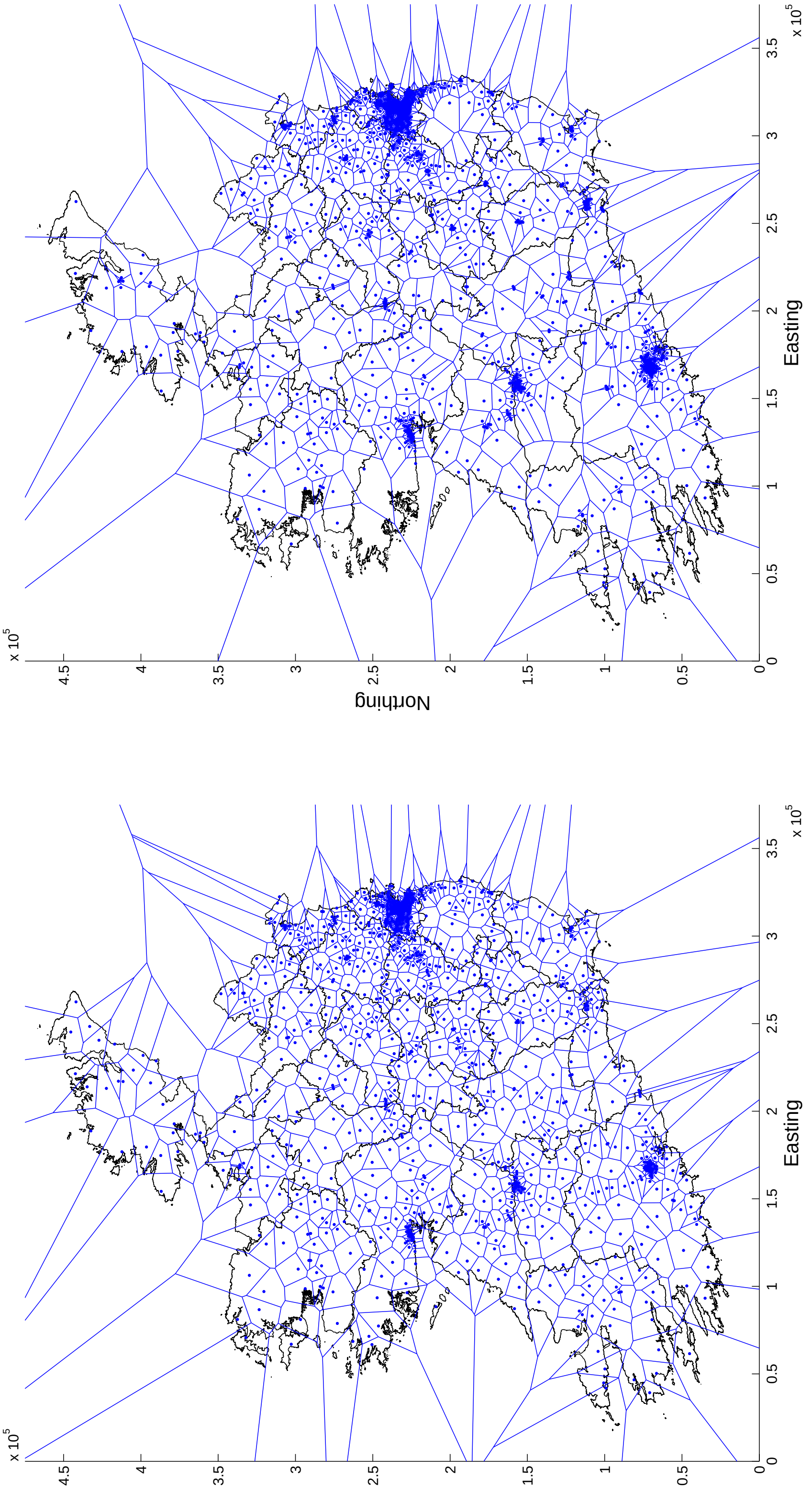
$$S_a = \frac{1}{2} \sum_{i=0}^{N-1} (x_i y_{i+1} - x_{i+1} y_i) \tag{3.2}$$

where N is the number of points in the coverage polygon and (x,y) are the spatial coordinates of each point. The effect of the limiting radius on the coverage map approximation is visualised in Figure 3.3. The limiting boundary is implemented by extracting the polygon of the spatial intersect of the idealised cell site coverage polygon with the polygon of the maximum cell site. This intersect is implemented using POLYBOOL, a function from the mapping toolbox of MATLAB®.

More specific cell sectored coverage regions may be extracted by incorporating the transmitter azimuth angle information into the tessellation, as visualised in Figure 3.4. This tessellation is achieved by subdividing each cell site coverage polygon by the unique transmitter azimuth angles of cells associated with the site. Note, cells with the same azimuth angles share the same cell coverage polygon, C_p. Cell radius (C_r) and area (C_a) may be calculated via equation 3.1 and equation 3.2 respectively. Also, an individual cell centroid Easting and Northing location, (C_x, C_y), may be calculated via equations (3.3) and (3.4), respectively.

$$C_x = \frac{1}{6C_a} \sum_{i=0}^{N-1} (x_i + x_{i+1}) (x_i y_{i+1} - x_{i+1} y_i) \tag{3.3}$$

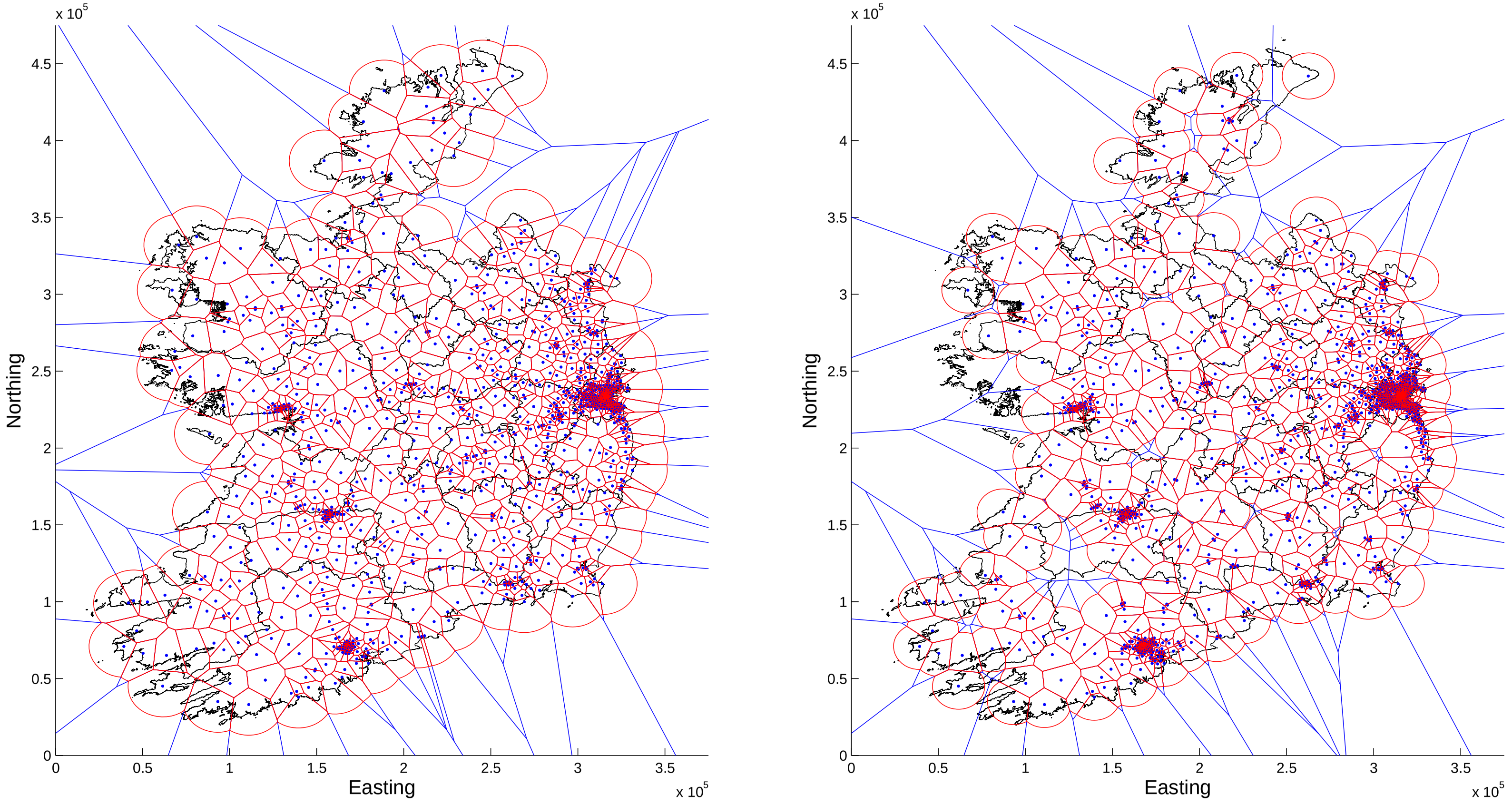
$$C_y = \frac{1}{6C_a} \sum_{i=0}^{N-1} (y_i + y_{i+1}) (x_i y_{i+1} - x_{i+1} y_i) \tag{3.4}$$



(a) 2G

(b) 3G

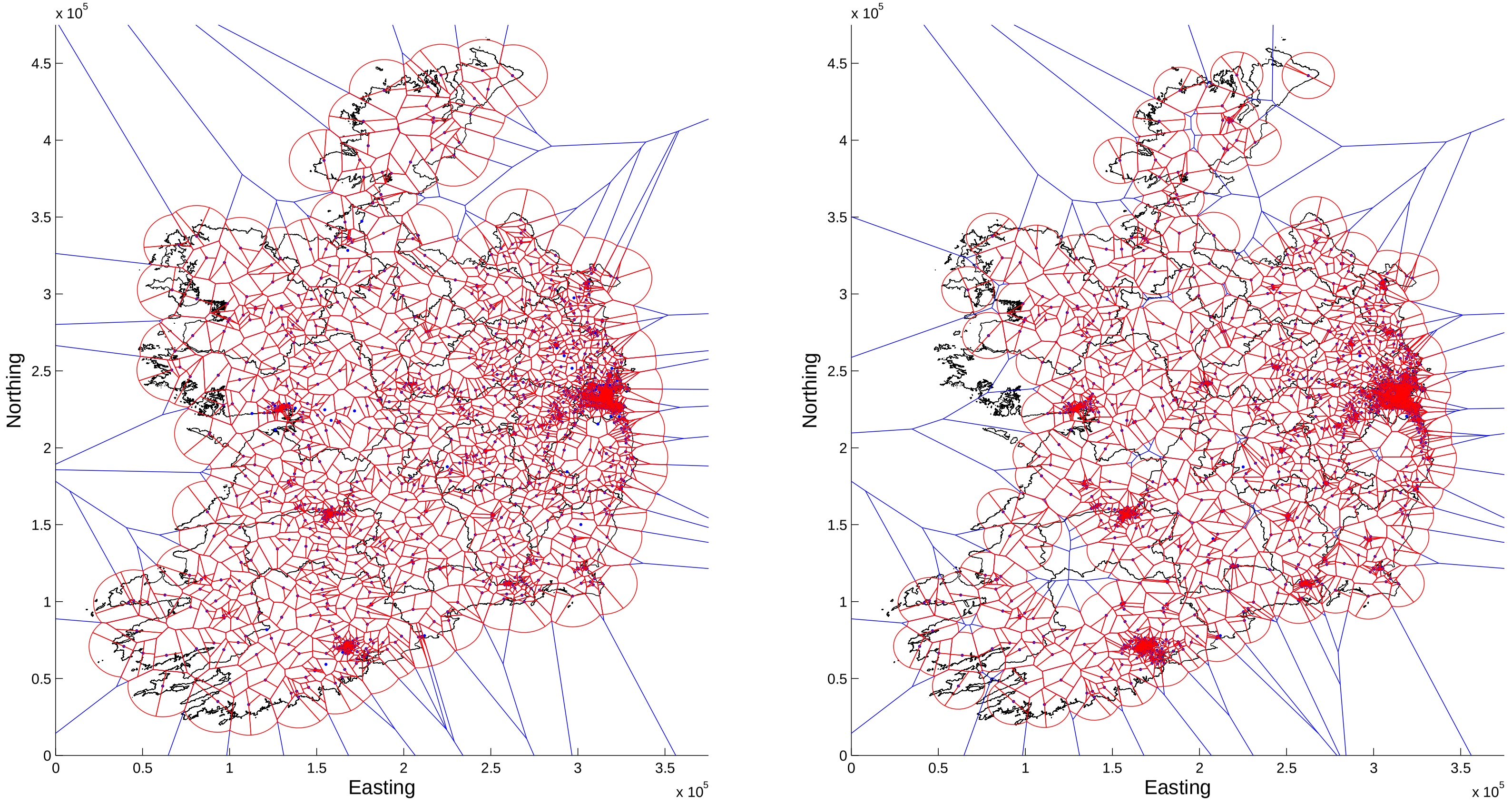
Figure 3.2: Voronoi diagram of 2G and 3G cell site coverage regions.



(a) 2G

(b) 3G

Figure 3.3: Restricted 2G and 3G cell site coverage regions.



(a) 2G

(b) 3G

Figure 3.4: Sectored 2G and 3G cell coverage regions.

3.2 Data Cleansing

It is common practice for mobile operators to relocate hardware within their mobile network. As a result, the spatial locations of cells may become outdated and can introduce errors to studies where location information is important. Such errors are clearly seen when the relationship between travel time and distanced travelled is observed. Due to localisation based on cell identification, there is uncertainty associated with CDR location estimates. If a subscriber is consecutively serviced by cells C_i and C_j at time t_i and t_j respectively, then the subscriber is assumed to have moved from C_i coverage polygon, C_{pi}, to C_j coverage polygon, C_{pj}, in time $t_j - t_i$ or less (i.e. $t_j - t_i$ is an upper bound on the time it took to travel from C_{pi} to C_{pj}). The uncertainty associated with the actual distanced travelled by the subscriber will be a function of cell topology. Using a cell's coverage polygon as an estimate of the possible locations, a subscriber may inhabit while being serviced by a particular cell, the distance travelled will vary between d_{ij}^{min} and d_{ij}^{max}. As depicted in Figure 3.5, the max distance between points in two polygon regions will always be two vertices so

$$d_{ij}^{max} = \max_{m,n} \|C_{pim} - C_{pjn}\| \tag{3.5}$$

where C_{pxy} is the y^{th} vertex of the cell coverage polygon for cell x, $m = [1 \rightarrow M_i]$, $n = [1 \rightarrow N_j]$, and M_i, N_j are the number of vertices in polygon regions C_{pi} and C_{pj}, respectively. However this is not always the case for the minimum distance as sometimes the shortest distance between two polygons could be between a vertex and a side. Therefore, d_{ij}^{min} is given by

$$d_{ij}^{min} = \min_{m,\gamma,n,\beta} \| [\gamma C_{pi,m} + (i - \gamma) C_{pi,m+1}] - [\beta C_{pj,n} + (1 - \beta) C_{pj,n+1}] \| \tag{3.6}$$

where $0 < \gamma < 1$, $0 < \beta < 1$, $m = [1 \rightarrow M_i]$, $n = [1 \rightarrow N_i]$ and $n = 1 + M_i \equiv 1$, $m = 1 + M_j \equiv 1$. The average distance, d_{ij}, can be estimated by taking the Euclidean distance between cell coverage polygon centroids.

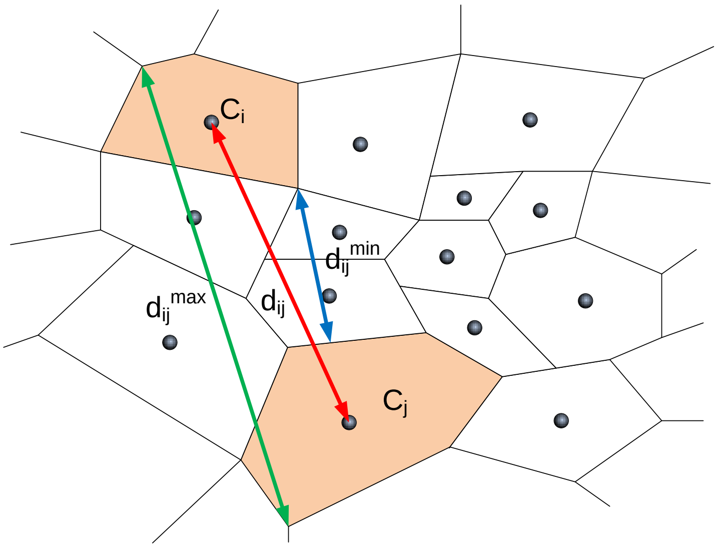


Figure 3.5: Depiction of the level of uncertainty associated with the distanced travelled between cells C_i and C_j from t_i to t_j. d_{ij}^{min} shows the shortest distance between the cells while d_{ij}^{max} show the longest distance between them.

As illustrated by Figure 3.6, plotting d_{ij} against the travel time t_{ij}, where $t_{ij} = t_j - t_i$, shows that there are occurrences of mobile subscribers being consecutively serviced by cells which are several hundred kilometres apart. Such observations do not adhere to practical travel speed restrictions. The baseline cell boundary error,

$$\max_{i,j\ adjacent} d_{ij}^{max} \tag{3.7}$$

is approximately 66.8 km from existing CDR data and is indicated by the green horizontal line in Figure 3.6.

As a result, a procedure is needed to identify those cells whose positional information has become outdated. If cells are adjacent then $d_{ij}^{min} = 0$, hence consecutive activities may produce $t_{ij} \approx 0$. As $t_{ij} \approx 0$ should only occur if cells are adjacent, the detection of $t_{ij} \approx 0$ for non-adjacent cells indicates that the location of C_i or C_j is outdated. However, cell coverage polygons are only estimates of cell spatial coverage, and as a result $t_{ij} \approx 0$ may occur in cells where C_i and C_j are relatively close. In general for two non-adjacent cells, the uncertainty associated with the actual distance travelled reduces if $d_{ij} \gg d_{ij}^{max} - d_{ij}^{min}$. Therefore, removing measurements where $d_{ij} \leq d_{ij}^{max} - d_{ij}^{min}$ reduces the error associated with determining outdated cells. Allowing for this, baseline cell boundary error and realistic travel speed expectations, a more realistic infeasible region can be defined. This region is illustrated in Figure 3.7.

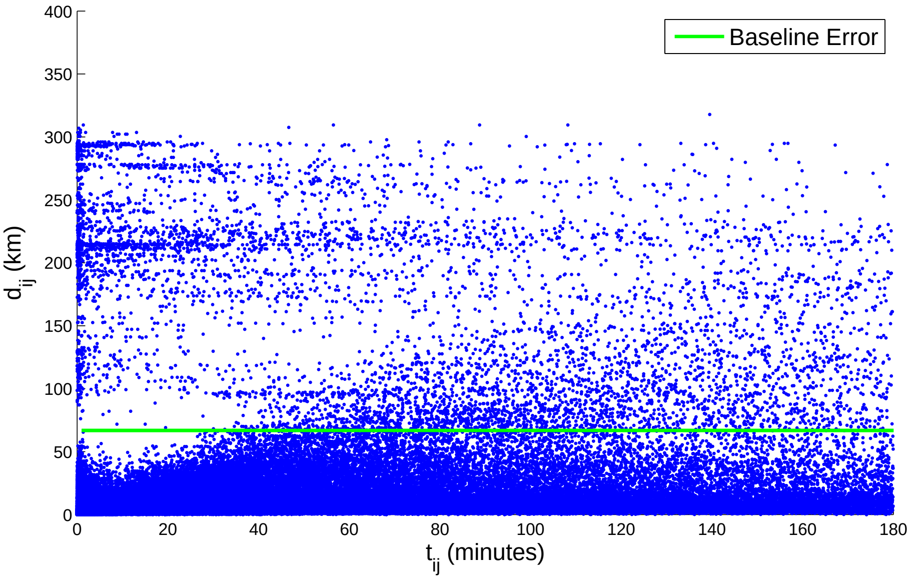


Figure 3.6: Distances travelled within a given time frame as observed through CDR.

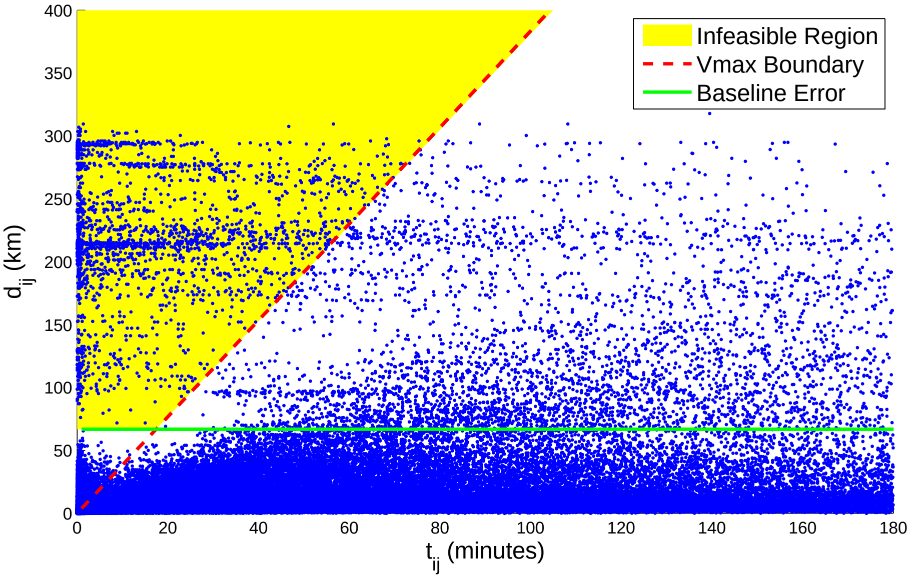


Figure 3.7: Distances travelled within a given time frame as observed through CDR with tolerance boundary and infeasible travel speed region superimposed.

If a data point (t_{ij}, d_{ij}) associated with a transition between C_i and C_j, falls within this region, then both cells involved in the movement are flagged as potentially outdated. By observing all other transitions between each flagged cell and other network cells, the outdated cell may be identified. To maintain consistency throughout our dataset, data from cells which have been identified as becoming outdated is removed. However in some instances, due to the low number of observed transitions, it may not be possible to isolate the offending cell. In this instance, both cells are marked as outdated. Figure 3.8 depicts d_{ij} against the upper bound travel time t_{ij} observed where all cell information identified as outdated is removed.

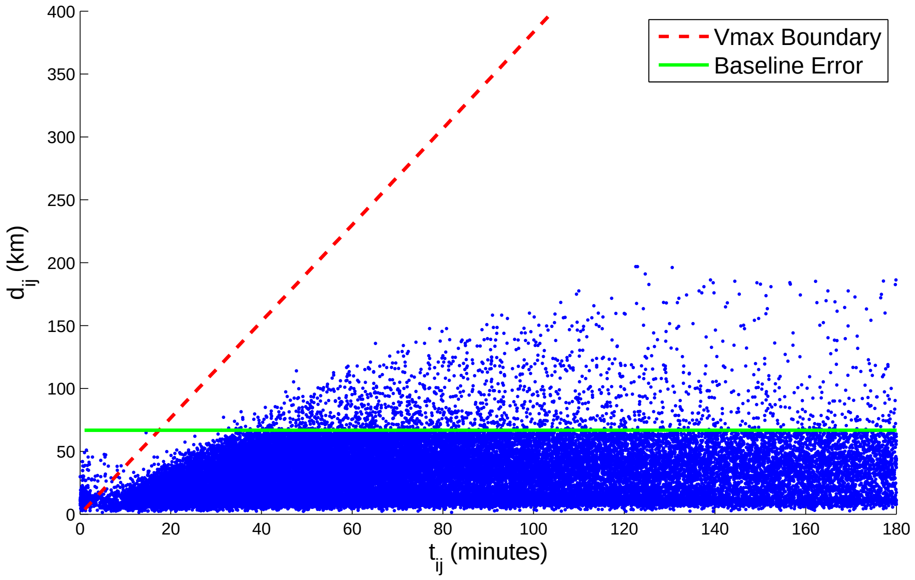


Figure 3.8: Distances travelled against the upper bound travel time as observed through CDR with applied filtering.

Due to the uncertainty in t_{ij} and d_{ij}, it is not possible to get reliable estimates for $velocity_{ij}$. However, if we plot d_{ij}^{min} against t_{ij} we can get a lower bound on the maximum travel velocity. Specifically the slope related to observed measurements (t_{ij}, d_{ij}^{min}) is a lower bound on the associated maximum travel velocity, $\hat{V}_{ij}$. This is because $d_{actual} \geq d_{ij}^{min}$ and $t_{actual} \leq t_{ij}$, therefore

$$\begin{gathered} \hat{V}_{ij} = \frac{d_{ij}^{min}}{t_{ij}} \\ velocity_{actual} \geq \hat{V}_{ij} \;. \end{gathered} \tag{3.8}$$

As illustrated by the manually drawn slope in Figure 3.9, a lower bound on the maximum $\hat{V}_{ij}$ (slope of black line) indicates that a journey between 40 km and 150 km will have a minimal maximum velocity of approximately 100 km/h, while journeys between 150 km and 200 km and will have a minimal velocity of 72 km/h respectively. These estimates correspond well to average speed measurements estimated by [175]. The decrease in average speed with increased distance relates to geographical constraints and the reduced opportunity for direct travel between journey end points as travel paths are restricted to the underlying transportation infrastructure.

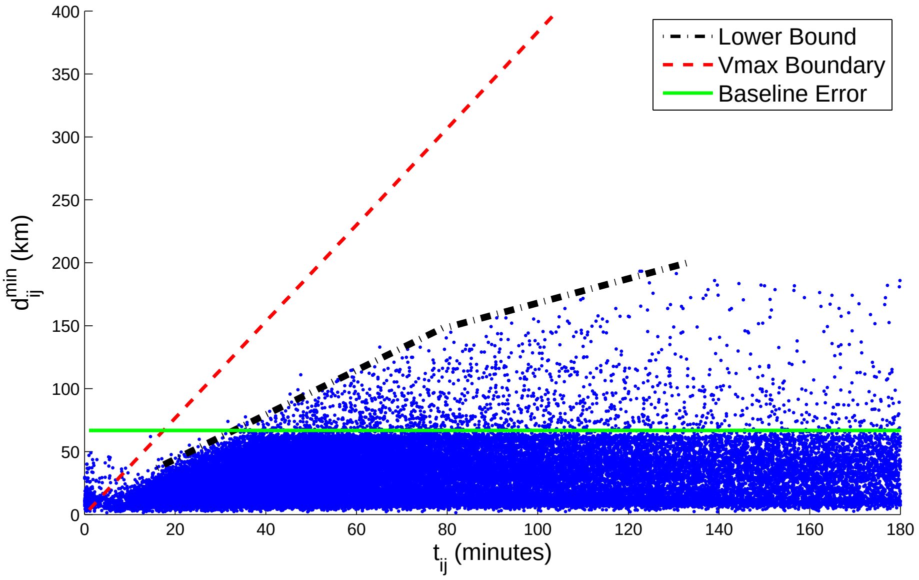


Figure 3.9: Distances travelled against the minimum travel time as observed through CDR with applied filtering.

3.3 Spatio-temporal Cell Activity Maps

Individual cell tower activities may be easily extracted from CDR tables by selecting temporally sorted rows which correspond to a cell of interest. A spatio-temporal cell activity map may then be constructed by combining cell tower spatial information, as outlined in section 3.1, with the temporally sequenced information extracted from CDR. For visualisation clarity, spatial smoothing of cellular activity across each cell coverage region is required. Standard temporal smoothing techniques may also be employed to remove high frequency temporal variations.

An example of temporal cell activity is depicted in Figure 3.10. Here, four cell towers were chosen at random from our mobile cell network, and their activities summed into 15 minute bins. Both 0 and 7 represent the midnight of Maonday (00:00:00). In order not to divulge possibly commercially sensitive information, each cell tower activity pattern is normalised. From this figure, the cycle of daily human activity is clearly evident, with expected lulls at night, and peaks during each day.

To enable a spatio-temporal cell activity map to be constructed from the temporal activity, a spatial smoothing function is required which accommodates varying cell size and multiple co-located and overlapping cell coverage polygons from both 2G and 3G networks. The spatial

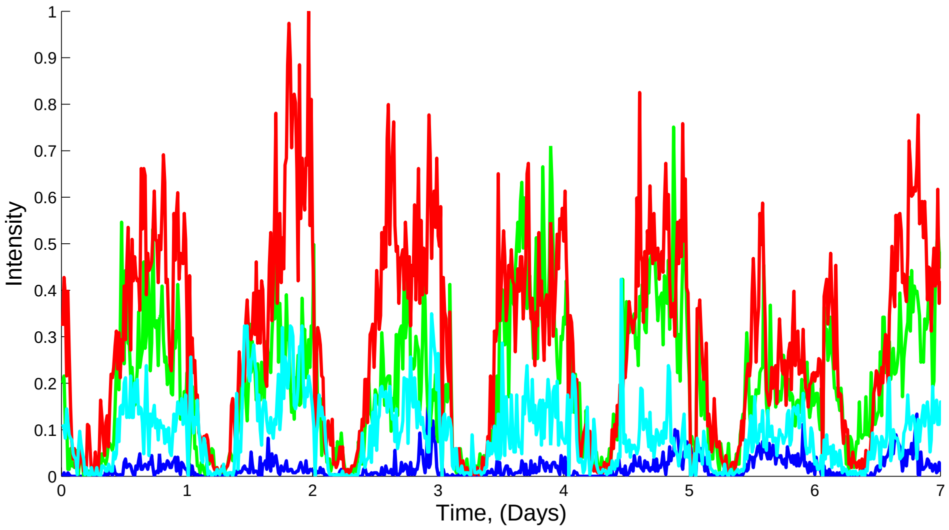


Figure 3.10: Temporal activity patterns of four randomly sampled cell towers over a seven-day period.

smoothing function involves the construction of an individual Gaussian bell for each cell tower, located at its coverage area centroid. The spreading factor of each Gaussian is governed by cell tower radius and spreads each individual cell metric of activity, C_w, over a spatial lattice, $\delta(x, y)$. An illustrative example of a weighted lattice for a single cell is depicted in Figure 3.11. The weighted spreading function of a single cell is given by

$$\delta(x, y) = \rho C_w \exp\left(-\frac{(x - C_x)^2}{2C_r^2} - \frac{(y - C_y)^2}{2C_r^2}\right), \tag{3.9}$$

where (x,y) are coordinates of points in the spatial lattice, C_r and (C_x,C_y) correspond to the cell radius and centroid location, respectively, and ρ is a scaling weight which ensures that the combined weights in $\delta(x, y)$ sum to C_w.

Each lattice, corresponding to an individual cell activity measurement may be expanded to the temporal horizon by incorporating an additional parameter at a particular time sample k. The resulting lattice $\delta(x, y, k)$ may then be combined to view the spatial distribution of activities at that instant. The combined weighted lattice, $\Phi(x, y, k)$, is given by

$$\Phi(x, y, k) = \sum_{C=1}^{N_c} \delta_C(x, y, k) \tag{3.10}$$

where $\delta_C(x, y, k)$ is the lattice for the cell tower C and N_c is the number of cell towers.

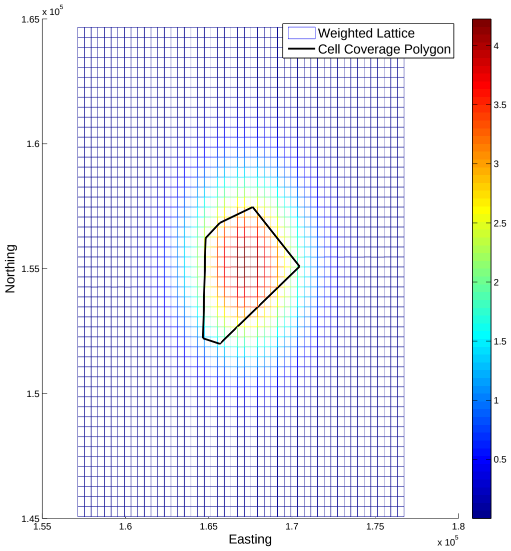


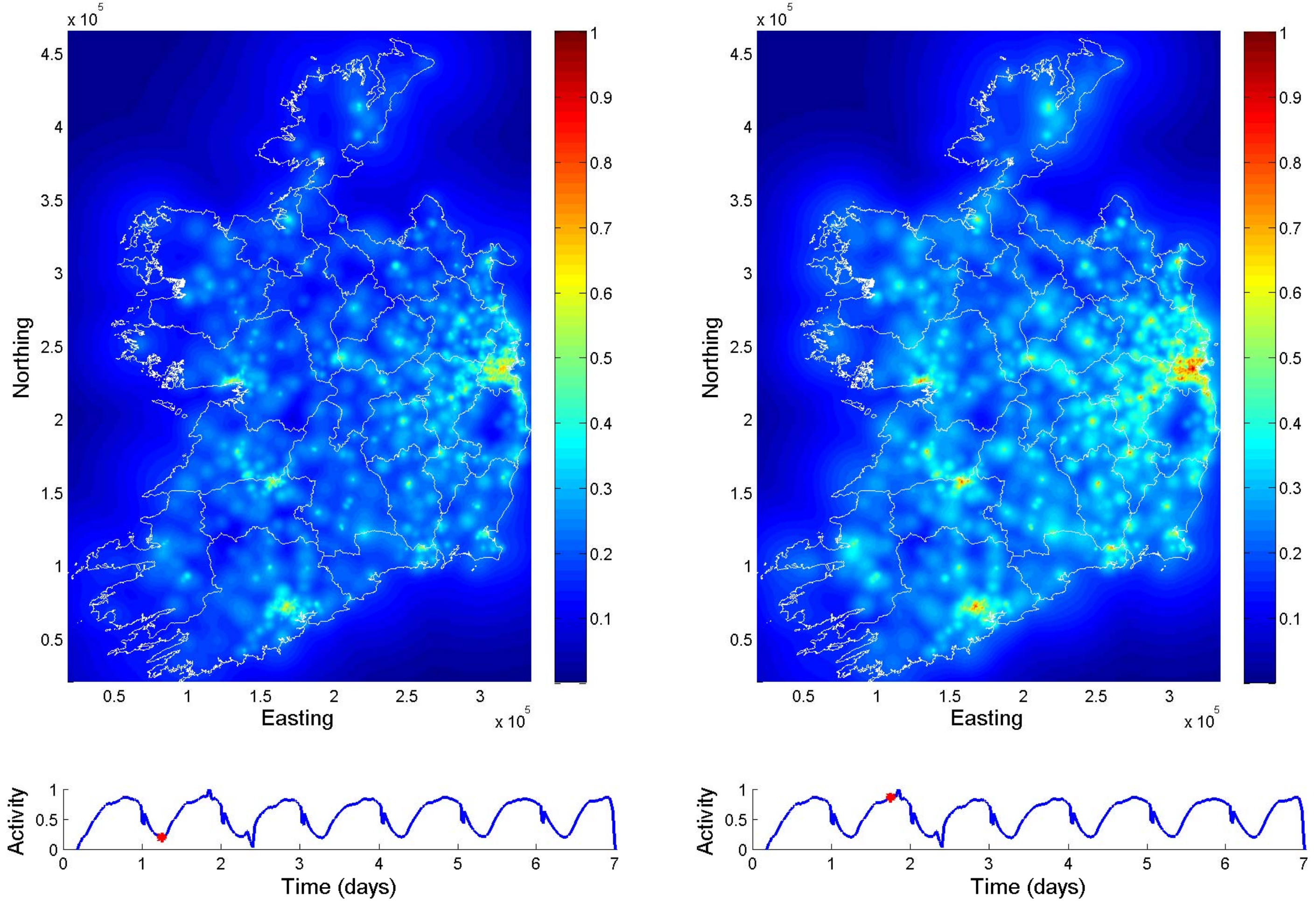
Figure 3.11: Example of a weighted lattice which corresponds to activity at a single cell tower. Here activity at a cell was set to 1000 and the cell radius is 2500 metres.

An example visualisation of the cell activity map for spatial distributions of data sessions, calls and SMS loads as recorded on 22/02/2011 for time periods of high and low activity are depicted in Figure 3.12, Figure 3.13 and Figure 3.14, respectively. To this end, the Republic of Ireland was divided into 500 × 500 metre pixels and traffic intensity was assigned at each pixel considering the aforementioned spatial distribution functions, while temporal activity load was binned into 15 minute time intervals. Each visualisation is constructed using MATLAB® plotting functions, with county geographical regions polygons sourced from Ordinance Survey Ireland (OSI). Total network temporal activity in each instance is also indicated on each figure and is normalised. The time sample corresponding to each spatial lattice is indicated on each temporal plot by '$*$'.

For visual clarity, high frequency temporal variations were removed using a temporal smoothing function. The function, given by Equation (3.11), is a moving average filter which averages the current measurement over three temporal samples.

$$\bar{\Phi}(x, y, k) = \frac{1}{3} \sum_{i=k-1}^{k+1} \Phi(x, y, i) \, . \tag{3.11}$$

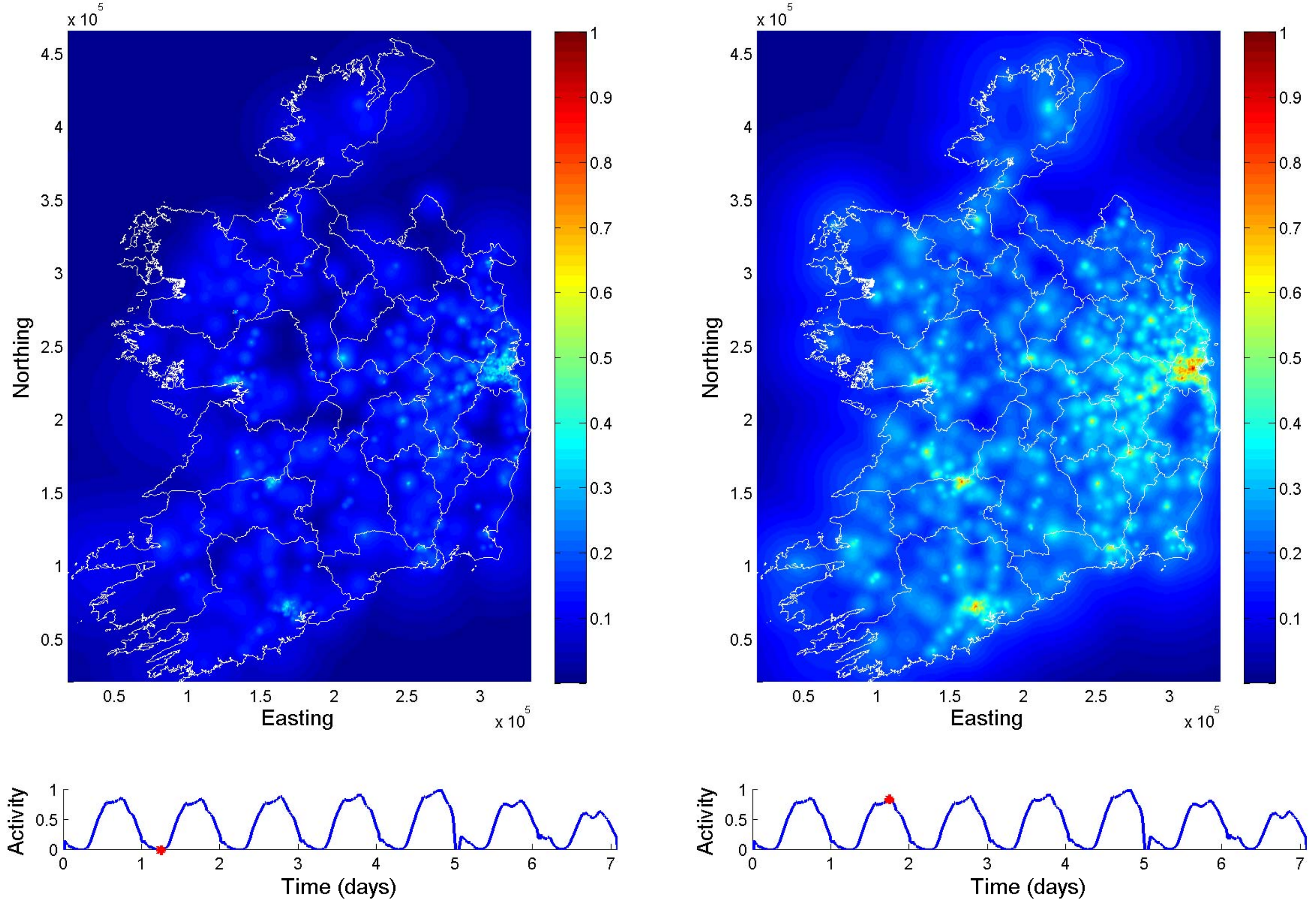
The plots show the dependence of activity levels with both population density and hour of the day. They also illustrate the relatively high spatial and temporal correlations of data, call and SMS loads.



(a) 22/02/2011 06:00:00

(b) 22/02/2011 20:00:00

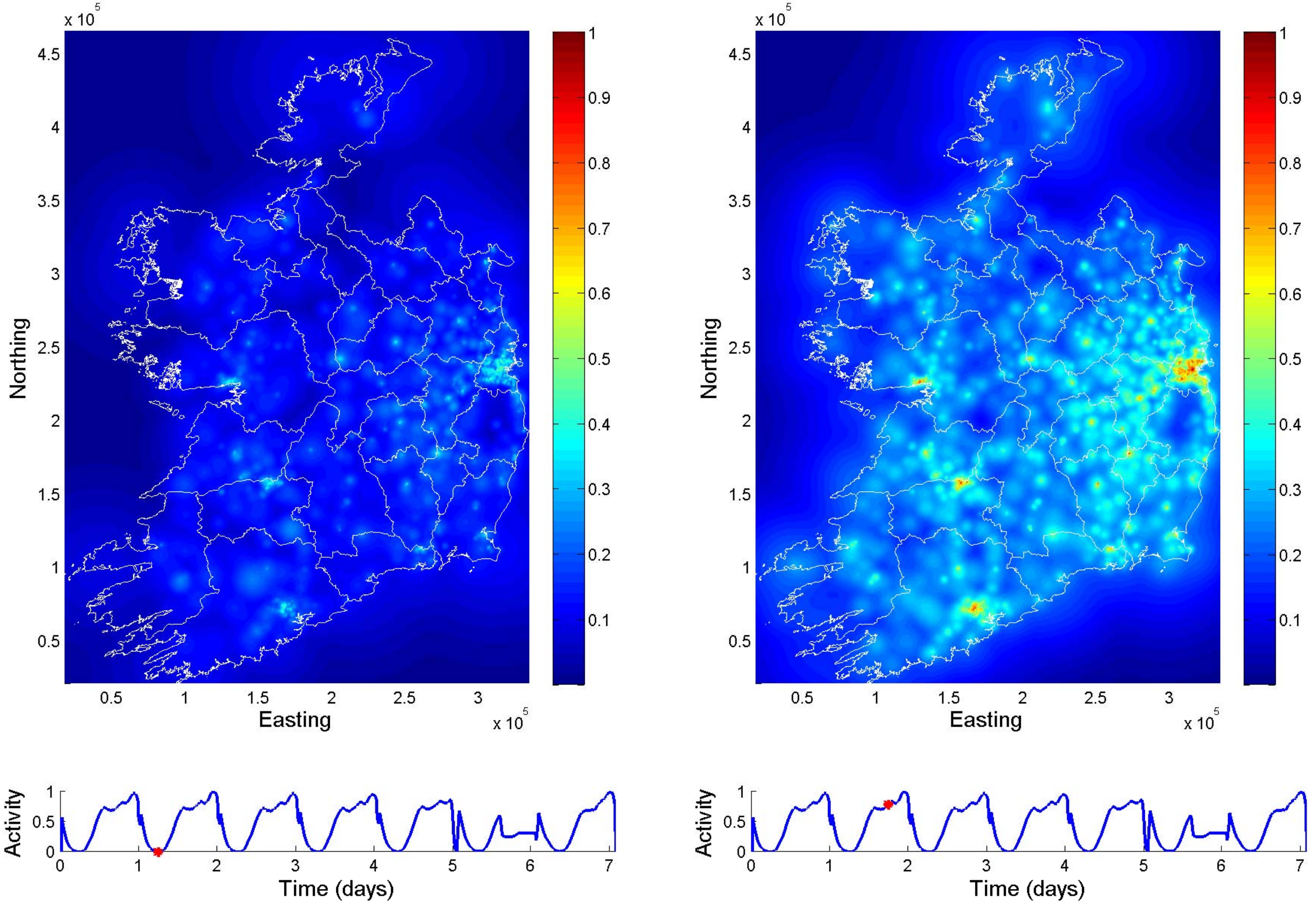
Figure 3.12: Spatial distributions of Data Sessions



(a) 22/02/2011 06:00:00

(b) 22/02/2011 20:00:00

Figure 3.13: Spatial distributions of Call Activity



(a) 22/02/2011 06:00:00

(b) 22/02/2011 20:00:00

Figure 3.14: Spatial distributions of SMS Activity

3.4 Mobile Device Trajectories

A mobile device CDR trajectory is the path that a subscriber follows through a cell network as a function of time as observed from CDR. Such trajectories can readily be extracted from CDR tables by selecting device specific temporally sorted cell tower connections. The trajectory may be spatially correlated by relating the spatial information of a servicing cell to each trajectory point. A sample of 5 randomly chosen observed trajectories is displayed in Figure 3.15, using a space time cube representation [176].

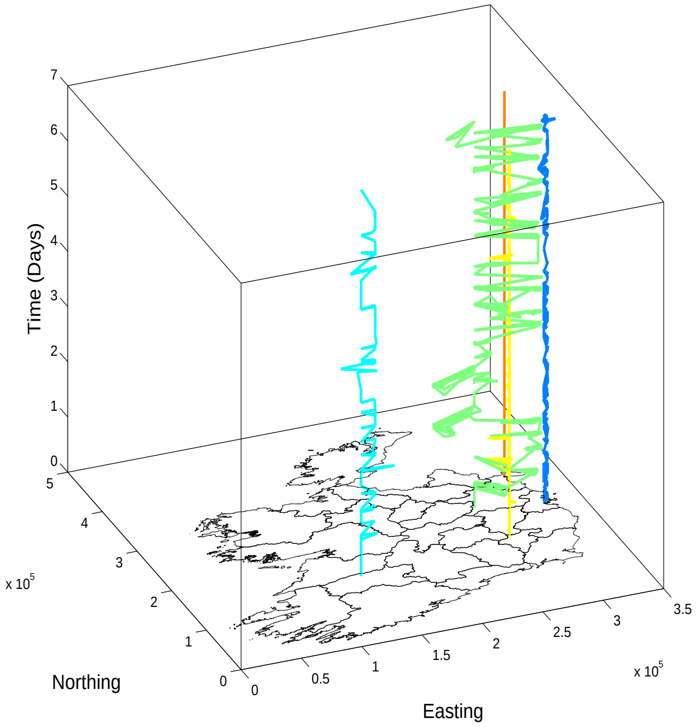
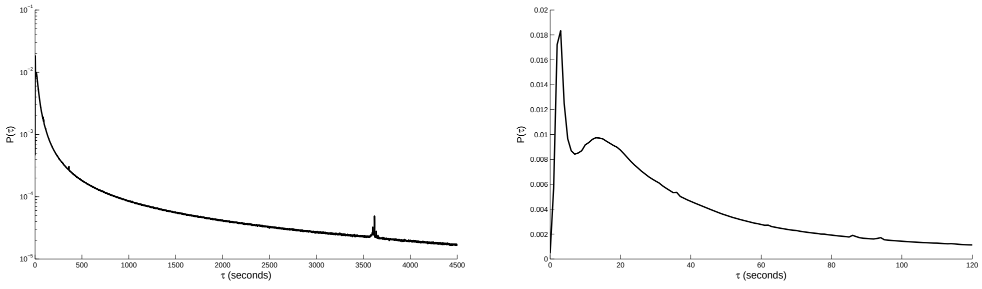


Figure 3.15: Space time cube visualisation of 50 user trajectories.

In general, location-specific information of trajectories generated in this way will be subject to spatial-heteroskedasticity, as the variance in estimation accuracy will be influenced by the variation in physical topology of the mobile network (i.e., the size and density of the cells). For example, it is less likely for a user 20 km away from a cell tower to be associated with that cell if the cell is located in the city centre compared to one located in rural areas.

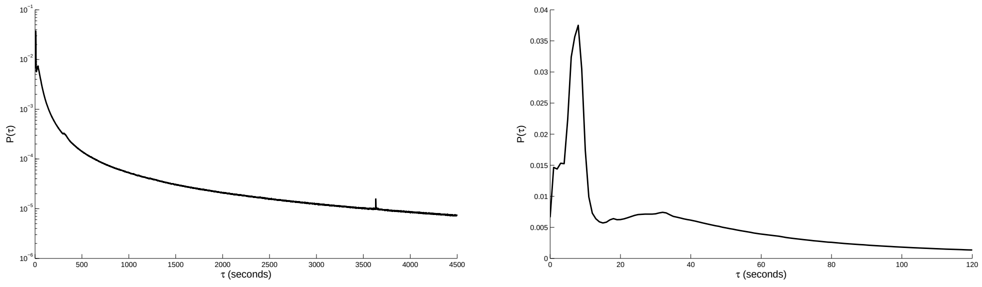
CDR trajectory sampling distributions, as discussed in [86], are non uniform and dictated by the activity profiles of individual subscribers. The distribution of the time intervals between consecutive activities, τ, across the whole operator's population under investigation for call, SMS and Data CDR are depicted in Figure 3.16, 3.17 and 3.18 respectively. Note that a bin size of one second is used to compute each distribution. As mobile devices can execute call, SMS and data sessions simultaneously, activity distributions may be combined when appropriate.



(a)

(b)

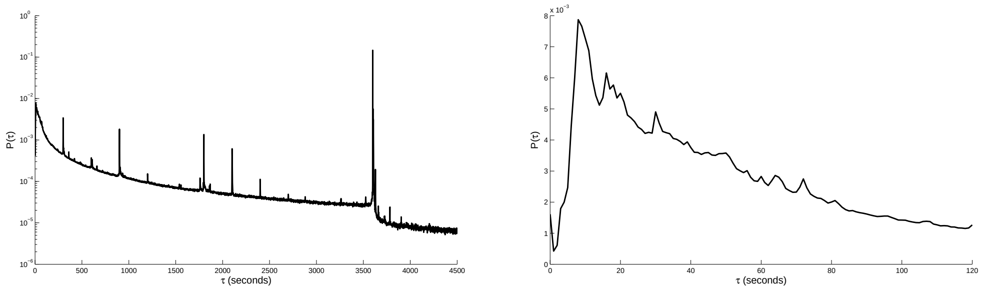
Figure 3.16: Probability density function (pdf) of time intervals between consecutive mobile calls: (a) over the range 0-4500 seconds; (b) over the range 0-120 seconds.



(a)

(b)

Figure 3.17: Probability density function (pdf) of time intervals between consecutive mobile SMS: (a) over the range 0-4500 seconds; (b) over the range 0-120 seconds.



(a)

(b)

Figure 3.18: Probability density function (pdf) of time intervals between consecutive mobile data sessions: (a) over the range 0-4500 seconds; (b) over the range 0-120 seconds.

From the pdfs, several temporal spikes in activity are noticeable. In each of the distributions, there exists a spike centred around the 1 hour mark. The 1 hour spike present in Figure 3.16a may be an artifact of Meteors billing policy, as under certain call plans Meteor only charges for calls which exceed an hour, thus encouraging customers to hang up and call again once that threshold approaches. This is supported by observing the 1 hour spike in the distribution of call durations, illustrated in Figure 3.19. The 1 hour spike present in Figure 3.17a may also correspond to network artifacts such as delivery of previously undelivered text message, or it may be as a result of periodic activities caused by automated systems which communicate over Meteor's mobile network. The noticeable weight corresponding the 1 hour spike in Figure 3.18a is linked to a procedure in Meteor's billing system which creates a new data session instance for sessions with duration over an hour. Observing the duration of data sessions in Figure 3.20, this procedure is clearly evident. The remaining spikes in Figure 3.18a may also correspond to periodic activities caused by automated systems which communicate over Meteor's mobile network, or may be artifacts introduced by applications commonly installed on mobile devices.

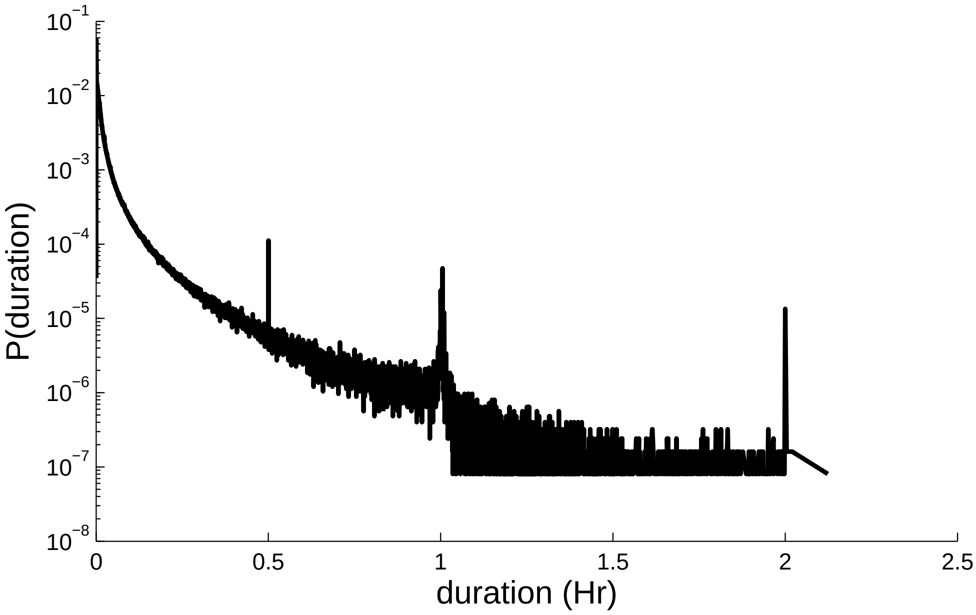


Figure 3.19: Probability density function (pdf) of the duration of calls.

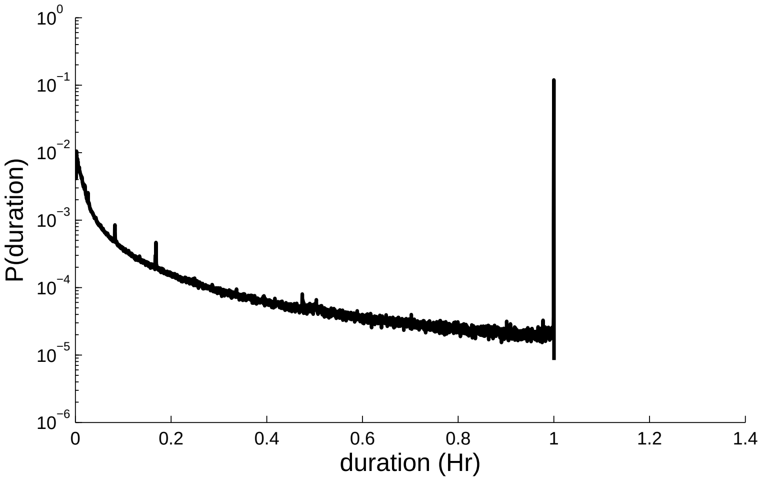


Figure 3.20: Probability density function (pdf) of the duration data sessions.

In Figures 3.16b and 3.17b, there are interesting double peaks clearly evident below the 60-second mark. In Figure 3.16b the peaks centred at 5 and 15 seconds respectively, may represent the observed process of making an unsuccessful call and attempting to redial. In Figure 3.17b, the double peaks are located at 10 and 30 seconds respectively. The initial sharp peak may refer to the process of sending a text message and receiving a subsequent delivery notification, while the second broader peak may reflect the average reply time.

The number of observed active subscribers over a 7-day period is illustrated in Figure 3.21. Note, similar to the cell tower activity distributions as discussed in Section 3.3 it is clearly evident that there exists a variability in trajectory sampling which is time dependant as it fluctuates with the cycle of daily human activity. Further insights into the temporal variability of CDR sampling is discussed by [64].

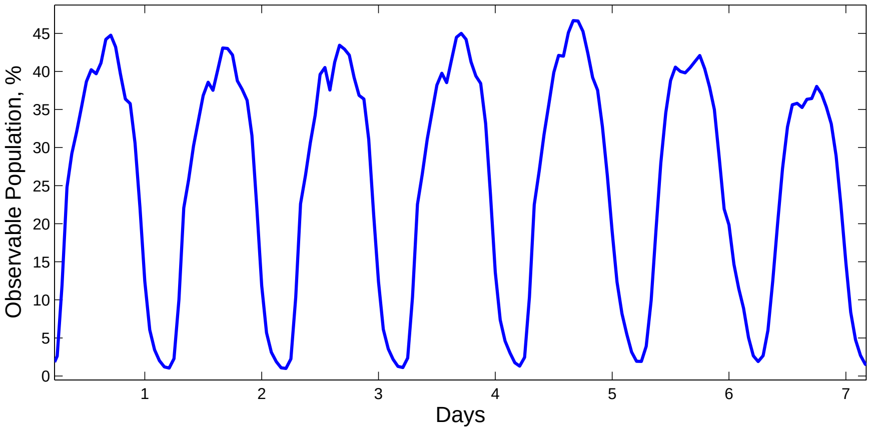


Figure 3.21: Proportion of observable user population over a seven-day period.

To investigate whether or not this temporal variation is statistically significant, the variance in expected τ, $E(\tau)$, over a 24-hour period was evaluated to see if the observed sampling distributions were indeed a function of time. Before the test of significance can be measured, a distribution of τ for each hour was tabulated, as depicted in Figure 3.22. Here, each observed value of τ captures the measured time between an activity that occurred within a temporal window and the next consecutive activity from an individual subscriber. From Figure 3.22, it is clear that none of the τ distribution are Gaussian. As a result, bootstrapping was applied to measure $E(\tau)$ form each temporal window.

Bootstrapping [177] is a resampling technique which consists of taking the mean of samples from a population with replacement in order to produce a distribution of mean values for that particular population. By central limit theorem, a distribution of mean estimates from samples of a population with finite variance approaches a normal distribution regardless of the statistical distribution of the population. As illustrated in Figure 3.23, the distributions of mean values for several temporal windows approach a normal distribution. Thus, the $E(\tau)$ for each temporal window can be approximated using the sample mean of each distribution to within a certain level of confidence. The variance in estimated $E(\tau)$ with 95% confidence intervals for each temporal windows is depicted in Figure 3.24. The variance in $E(\tau)$ and number of samples within each temporal distribution indicate that, on average, the sampling rate of active subscribers is less during the early hours of the morning.

During the hypothesis testing Welch's t-test [178] is used to evaluate if distributions of τ were sourced from distributions with statistically similar means and variance. While the t-test is not valid for small samples (N<30) from non-Gaussian distributions, it is valid for large samples. This is because of the fact that it makes no assumptions on the normality of the distribution, rather is assumes that the mean value of $E(\tau)$ is normally distributed. Applying Welch's t-test [178] to each pair of observed τ distributions we find that each distribution of τ is not statistically similar to all other samples at 95% significance. As a result, the aforementioned sampling distributions are temporally heterogeneous. Accordingly, observed journeys taken during times of low network activity will be generally undersampled compared to journeys taken during times of high network activity due to increase in $E(\tau)$.

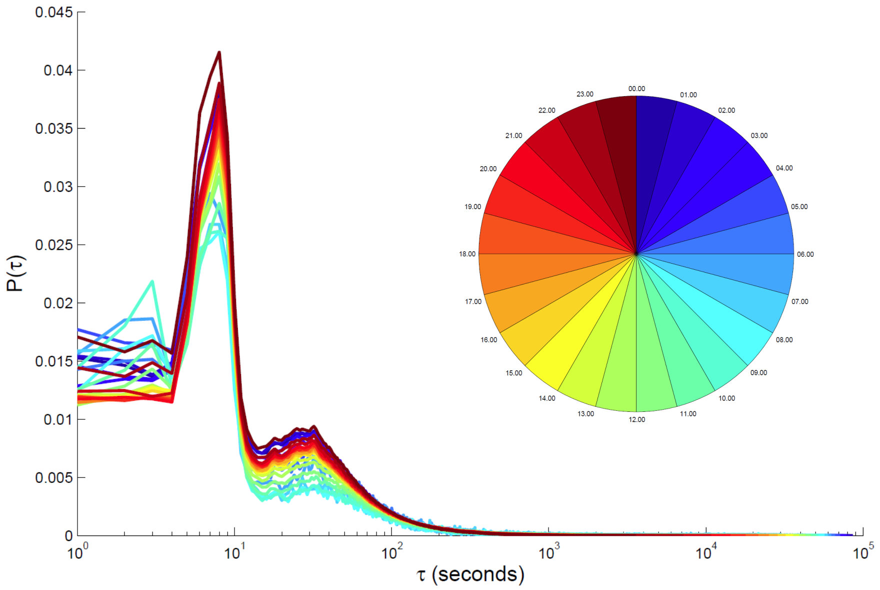
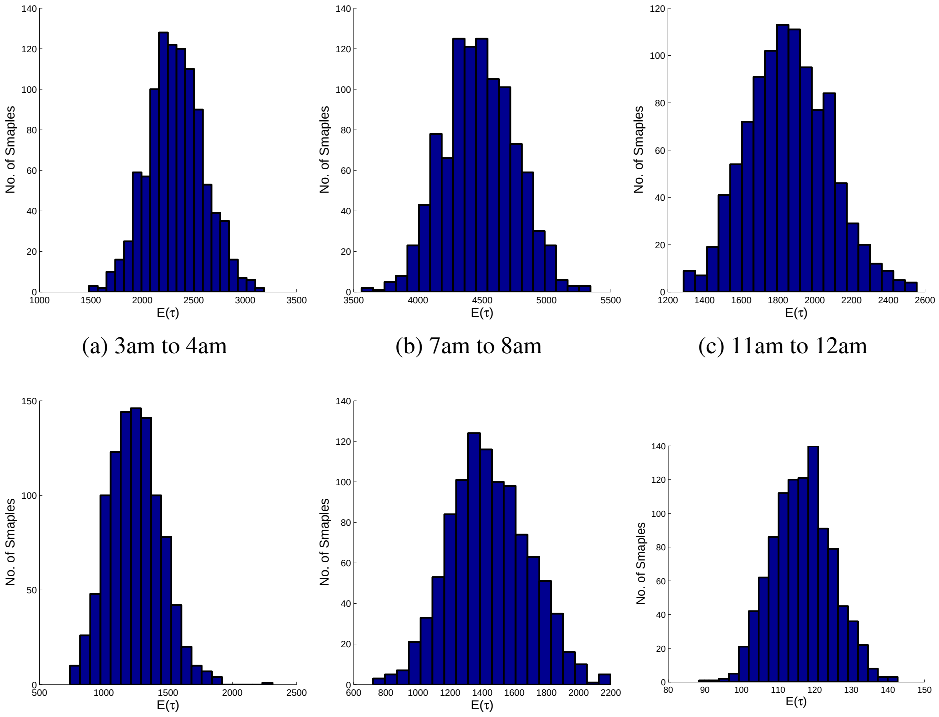


Figure 3.22: τ distributions observed over a 24-hour period, where each distribution of τ captures the observed times between activities which occurred within a temporal window and the next consecutive activity for individual subscribers. Temporal information is encoded using the colour chart provided.



(a) 3am to 4am

(b) 7am to 8am

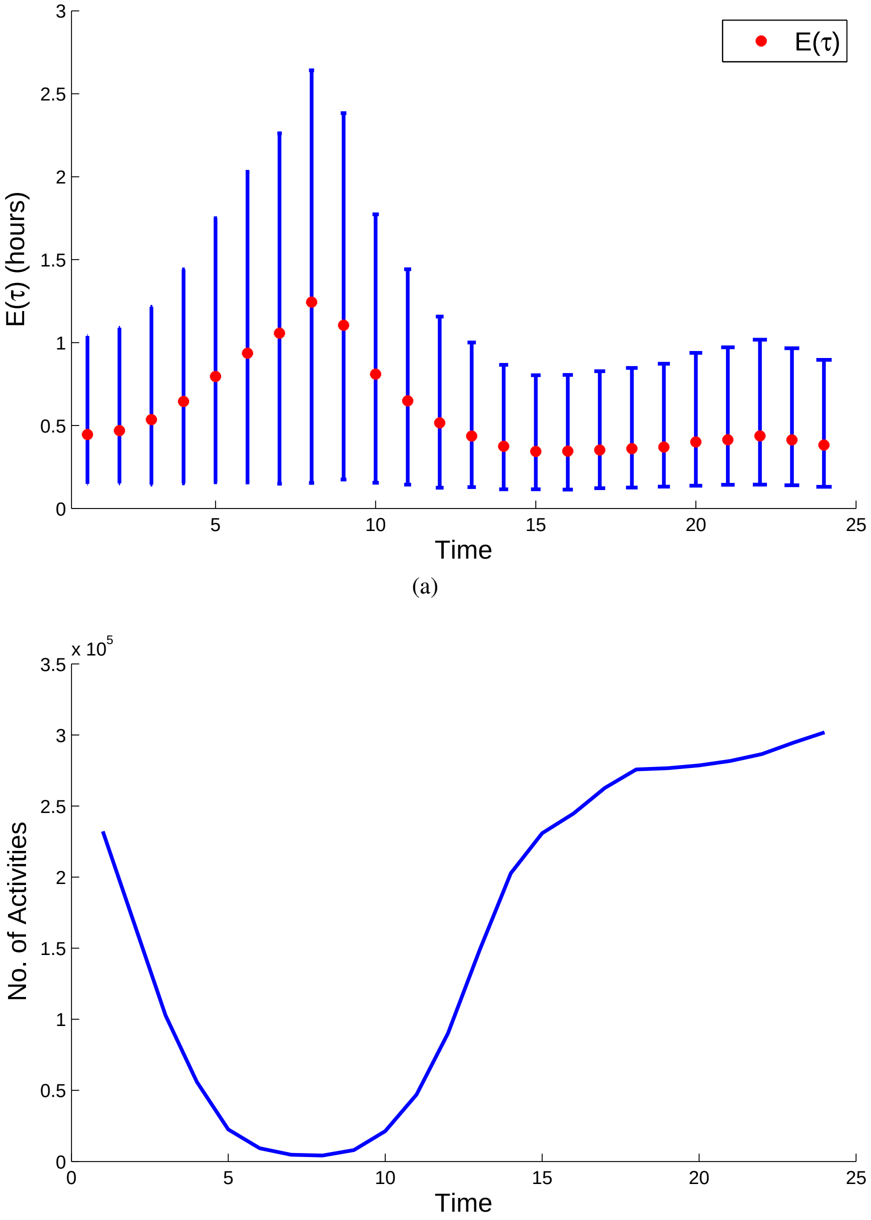
(c) 11am to 12am

(d) 3pm to 4pm

(e) 7pm to 8pm

(f) 11pm to 12pm

Figure 3.23: Distributions of the mean time to the next subscriber activity, where the initial activity occurred at; (a) 3am to 4am; (b) 7am to 8am; (c) 11am to 12am; (d) 3pm to 4pm; (e) 7pm to 8pm; and (f) 11pm to 12pm.



(a)

(b)

Figure 3.24: The variance in $E(\tau)$ over a 24-hour period: (a) $E(\tau)$ with 95% confidence intervals; and (b) the number of activities.

3.5 Spatial Clustering of Cell Towers

A mobile network topology is governed by coverage and capacity requirements. While cell coverage is generally influenced by geographical factors, capacity is generally influenced by traffic demand [1, 65, 93, 94, 95]. As traffic demand is strongly linked to user population density, cell size and density vary with mobile user density. As previously discussed, typically mobile network topology for 2G, 3G and 4G are designed separately. This results in several cells of varying standard covering a single geographical area.

Several clustering methods may be used to combine multiple cell coverage polygons of varying standard into a single polygon representative of a symbolic location covering a population centre. To this end, an agglomerative hierarchical based clustering algorithm [179] was applied. The clustering similarity metric was the Euclidean distance between each cell's site location, including both 2G-3G distance measurements for our network of interest. The cell location information was inputted into the MATLAB® function LINKAGE, which returned a hierarchical cluster tree. The dendrogram of this tree is illustrated in Figure 3.25. Clustering was implemented on this tree using the MATLAB® function CLUSTER.

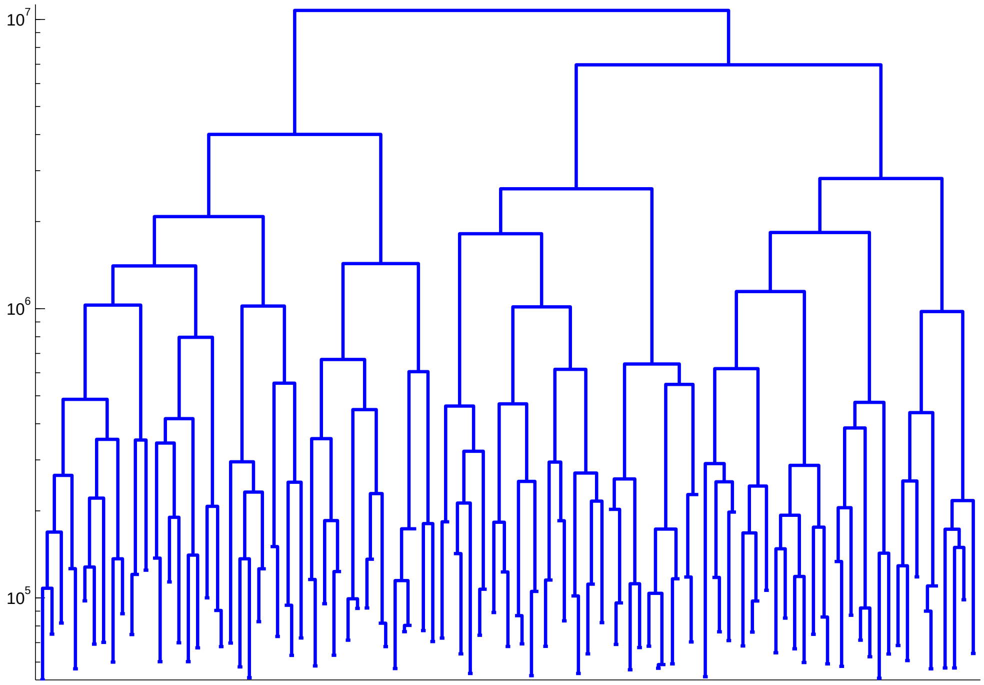


Figure 3.25: Dendrogram of hierarchical cluster tree. Note for visual clarity, the number of leaf nodes is limited to 200. The full tree has over 10,000 nodes.

This resulted in 500 distinct clusters with cells being grouped together if they were in close spatial proximity. By performing a spatial union on the coverage polygons (Section 3.1) of individual cells within each cluster, that cluster can be characterised by its regional coverage polygon. The visualisation of the clustered polygons is depicted in Figure 3.26.

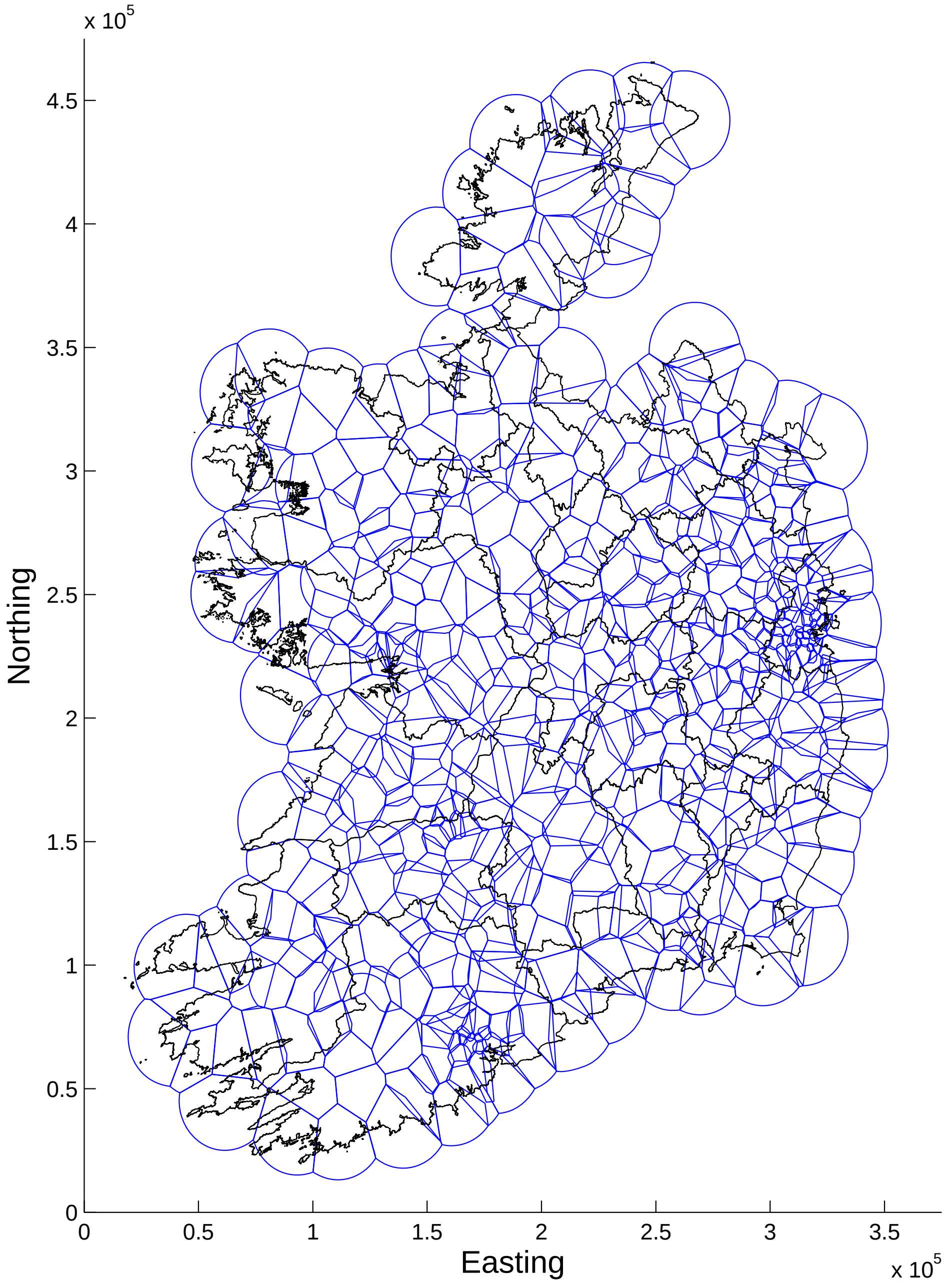


Figure 3.26: Regional coverage polygons.

3.6 Movement Transition Flows

With the ever increasing availability of trajectory data, observing the aggregated flows of people or animals between regions of interest has been a growing area of research. The work of Natalia and Gennady Andrienko *et al.* [27, 180, 181], Buchin *et al.* [182] and Doyle *et al.* [148] have explored varying techniques to visualise and group similar movement patterns. Similarly, CDR subscriber trajectories may be exploited in this regard.

By counting the number of subscriber transitions between servicing cell towers in a given time frame, we can construct an aggregated transition matrix, $\Upsilon_a(k)$,

$$\Upsilon_a(k) = \begin{pmatrix} \upsilon_{1,1}(k) & \upsilon_{1,2}(k) & \cdots & \upsilon_{1,N_R}(k) \\ \upsilon_{2,1}(k) & \upsilon_{2,2}(k) & \cdots & \upsilon_{2,N_R}(k) \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{N_R,1}(k) & \upsilon_{N_R,2}(k) & \cdots & \upsilon_{N_R N_R}(k) \end{pmatrix} \tag{3.12}$$

where N_R is the number of regions of interest, and $\upsilon_{ij}(k)$ is the transition intensity from region i to j at time k. For the mobile network considered in this research, Υ_a is a large matrix containing close to 115 million elements (N_R = 10,721). For transition flow analysis, the matrix size needs to be reduced in order to lower both the computational complexity and memory requirements.

Initially, as a compromise between computational load and spatial accuracy, 2G and 3G network cell towers were combined into 500 clustered cell regions as outlined in Section 3.5. This reduces Υ_a to Υ (N_R = 500). The flow of people between clustered regions and the geographical areas covered represents a proxy for the flow of people between individual population centres. The proportional link strengths demonstrating observed transitions between regions is illustrated in Figure 3.27.

The transition intensity or strength can also be observed temporally between two individual regions. A comparison of average daily activity volumes between Maynooth and Leixlip, towns in the north-east corner of County Kildare, Ireland, is displayed in Figures 3.28 and 3.29. With populations of 12,510 and 15,452, respectively, both are served by a commuter train service to Dublin and are close to the M4 motorway. From the figure, as expected Friday evening (4 - 8pm) has the highest transition volume in comparison to other days of the week. The commuting behaviour that exists between Maynooth and Leixlip is also evident, as early spikes in transition intensity from Maynooth to Leixlip is recorded on week days, with the expected lull on weekends.

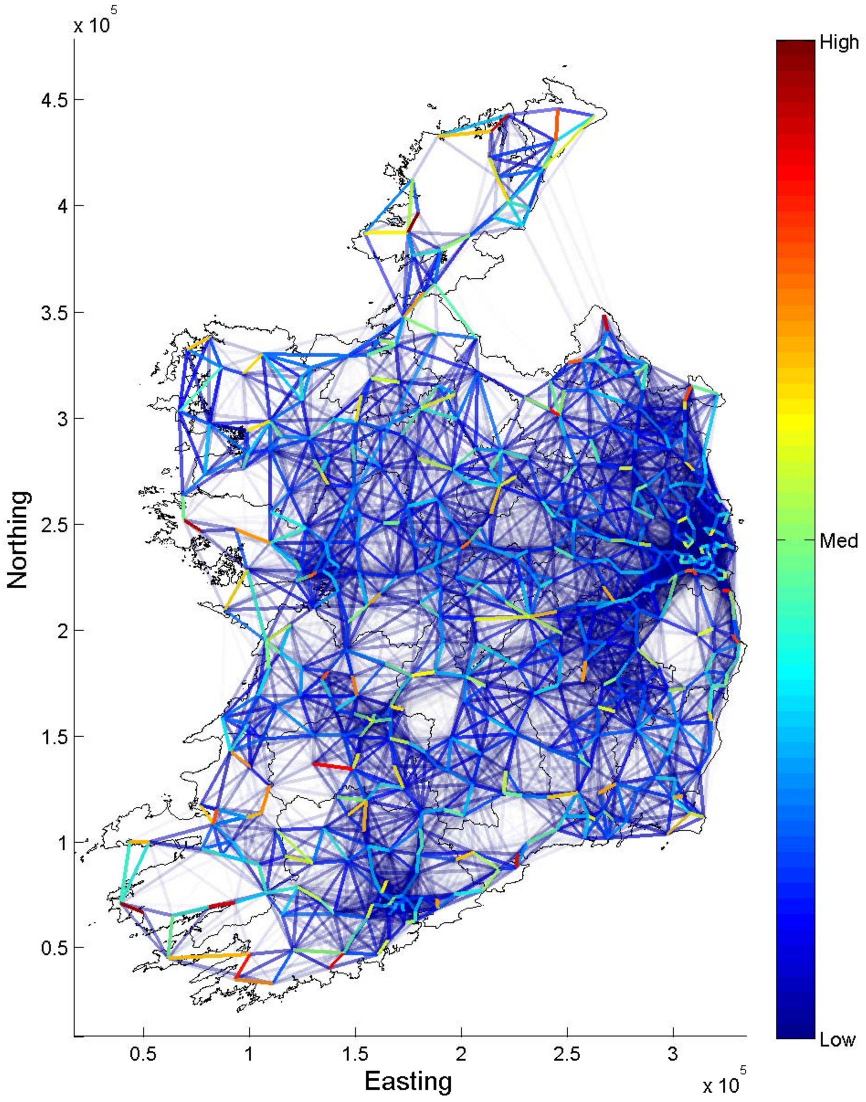


Figure 3.27: The proportional link strengths demonstrating observed transitions between clustered cell regions.

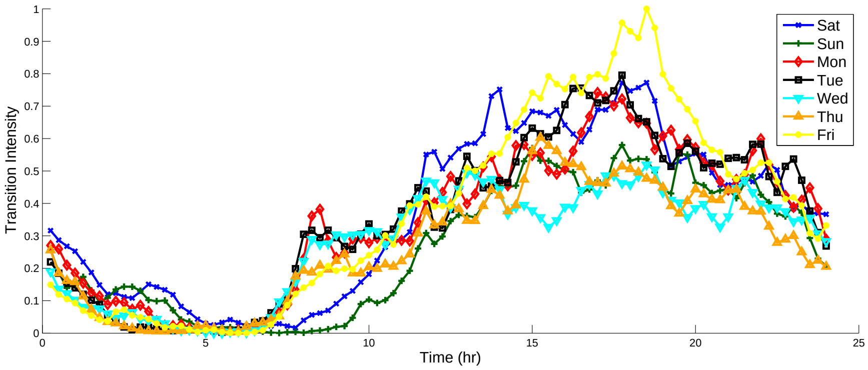


Figure 3.28: Average daily activity volumes of subscribers moving between clustered regions covering the towns of Maynooth and Leixlip, in the direction of Maynooth to Leixlip.

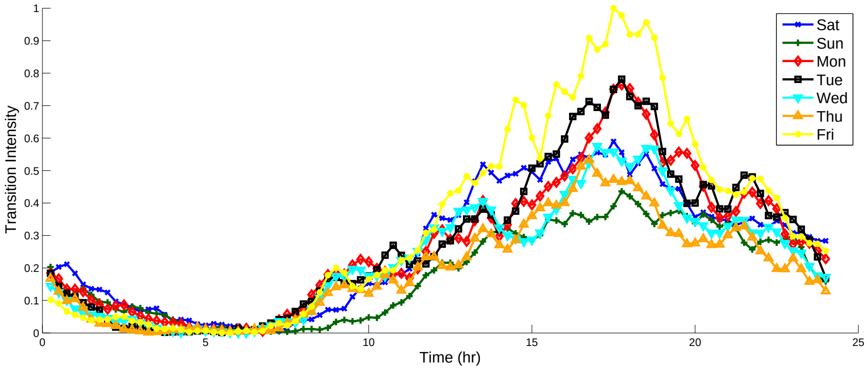


Figure 3.29: Average daily activity volumes of subscribers moving between clustered regions covering the towns of Maynooth and Leixlip, in the direction of Leixlip to Maynooth.

3.7 Communication Flow

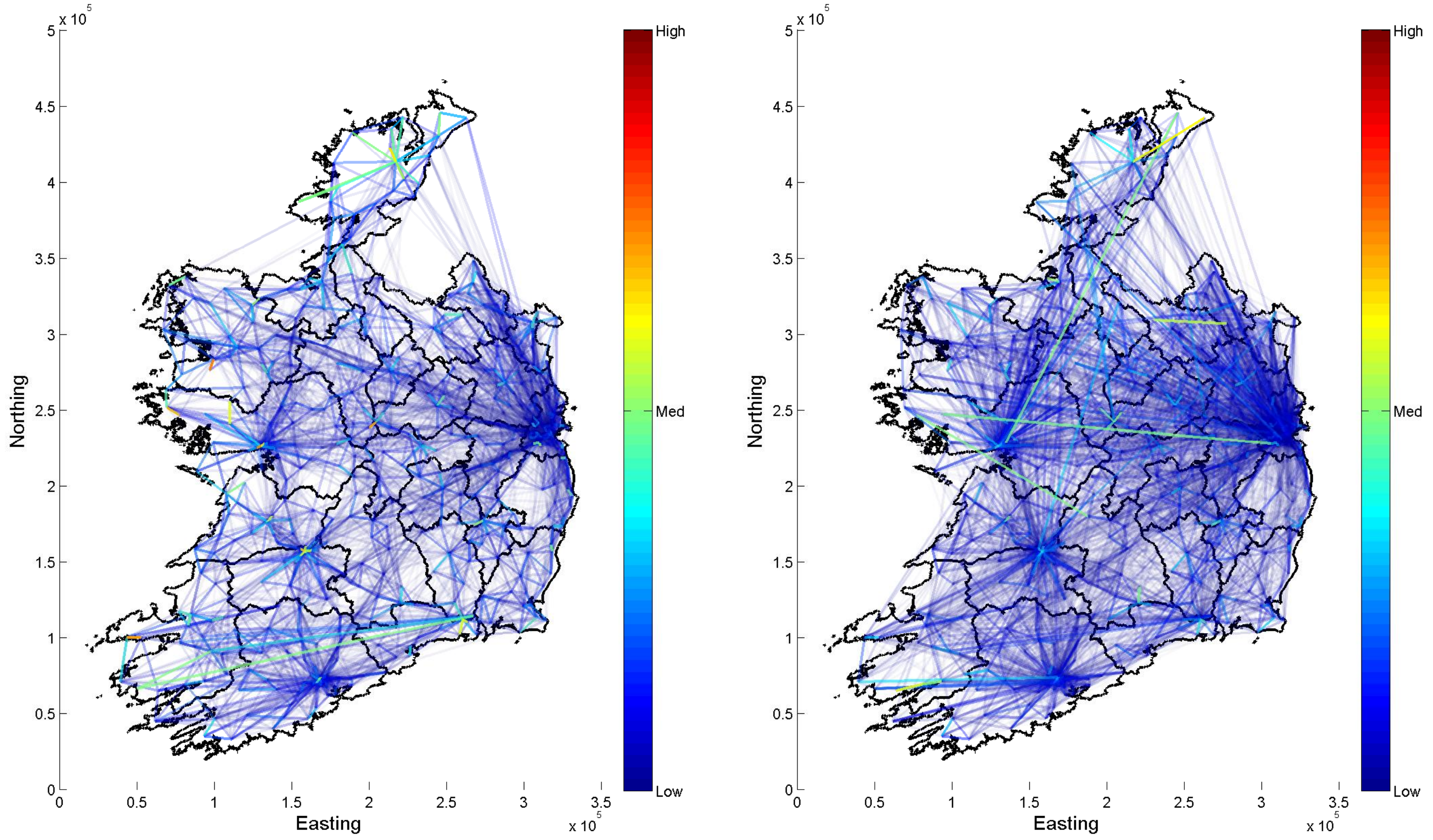
The flow of information from one region to the next is naturally observed through analysing mobile telephony data. As such, it has been the subject of many research articles in recent years [129, 161, 162, 163, 164]. In a similar fashion to the transition matrix discussed in Section 3.6, communication flow may be summarised by an aggregated communication flow matrix $\zeta_a(k)$,

$$\zeta_a(k) = \begin{pmatrix} \varsigma_{1,1}(k) & \varsigma_{1,2}(k) & \cdots & \varsigma_{1,N_R}(k) \\ \varsigma_{2,1}(k) & \varsigma_{2,2}(k) & \cdots & \varsigma_{2,N_R}(k) \\ \vdots & \vdots & \ddots & \vdots \\ \varsigma_{N_R,1}(k) & \varsigma_{N_R,2}(k) & \cdots & \varsigma_{N_R N_R}(k) \end{pmatrix} \tag{3.13}$$

where $\varsigma_{ij}(k)$ is the intensity of communication between the ith and jth region at time k. A sample of communication flow between regions for call and SMS is depicted in Figure 3.30. For visual clarity, $\zeta(k)_a$ in each instance has been modified such that $\varsigma_{ii} = 0 \; \forall \; i$. ζ_a is then normalised such that for any row i,

$$\zeta_a(k) = [\varsigma_{ij}(k)]_{N_R \times N_R} \rightarrow \sum_{j=1}^{N_R} \varsigma_{ij}(k) = 1 \; , \; \forall \; i \tag{3.14}$$

Each visualisation is constructed using MATLAB® plotting functions. Note that the opacity of each observed connection edge is dictated by its communication intensity value.



(a) Call

(b) SMS

Figure 3.30: Communication flow between clustered cell regions: (a) Calls; and (b) SMS.

3.8 Subscriber Social Graph

Typically, CDR contain information on both the sender and recipient of each voice or SMS communication. In such cases, a social graph may be constructed from the observed communications, where subscriber identifiers are used for the graph vertices, and communication intensity for graph edges [90, 159].

For the mobile network considered in this research, a complete social graph for Meteor to Meteor subscriber connections may be constructed. Partial information also exists for connections to subscribers on external operators through their interaction with Meteor customers. A sample social graph from a selection of Meteor's subscriber base is depicted in Figure 3.31. This graph is visualised using ORA® [183], a network visualisation software platform. Subscribers are orientated using a spring embedded layout [184]. Such graphs provide a rich source of information for research into social interaction and reaction. However, the analysis of user connections is not the focus of the research in this thesis and as such it is only mentioned here for completeness.

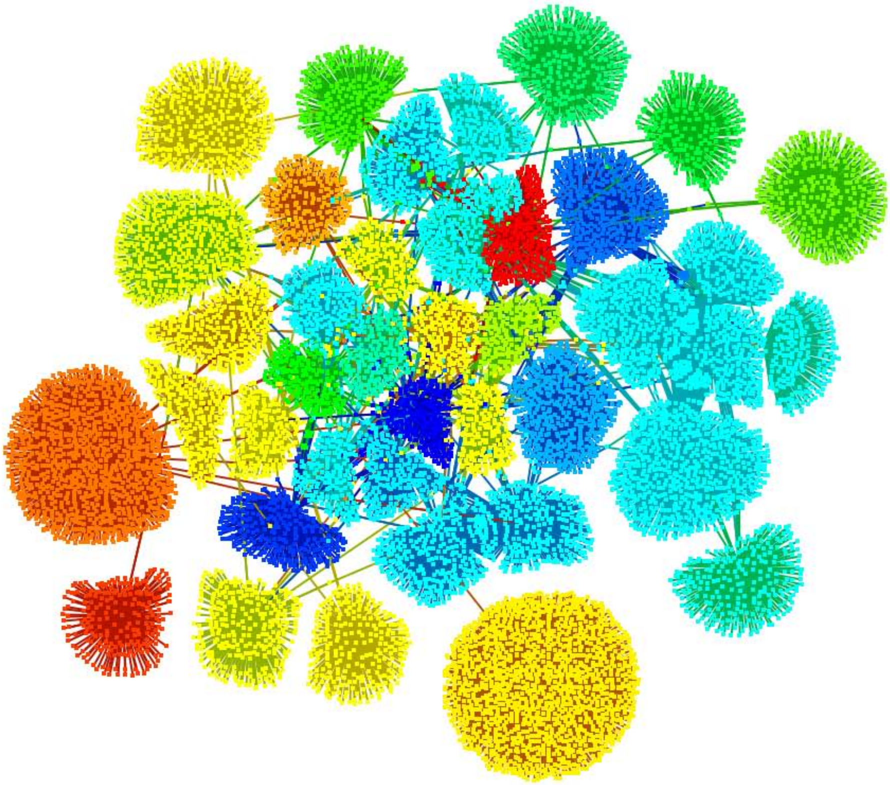


Figure 3.31: Sample user connection graph.

3.9 Discussion

This chapter presented the methodology used to extract and visualise a range of features from call detail records (CDR), cell tower information and subscriber registration data from the Meteor cellular phone network in the Republic of Ireland. Insights into required procedures for data cleansing, cell coverage area modelling and cell clustering were also presented. The contributions of the chapter include the estimation of the achievable distance a mobile device may travel over time, a novel activity spatial weighting function for spatio-temporal cell activities and a detailed discussion on the time variability of CDR trajectory sampling.

There are many potential applications which require a human activity feed or source, whether live or historic, to help explore useful real-time and predictive human-related behaviours. A number of the features presented in this chapter may be appropriate as inputs to such applications, the obvious example being subscriber CDR trajectories. Subscriber trajectories exhibit repetition as people often perform periodic journeys. By modelling a subscriber's movement patterns, we can predict their location in time. With such information it is possible to develop population density estimators, identify city and regional catchment areas, produce traffic monitoring applications and observe how urban environments are used. As the capacity to predict population density over time evolves, such information may be related to the demand loads seen on utility networks, thus enabling the development of load forecasting applications.

Dynamic transportation services may also benefit from such data. As the daily movement of people becomes more predictable, the location and quantity of public transportation needed to service a geographical area can be refined to meet current requirements. Also, real time traffic information can be relayed to traffic management systems enabling the real-time optimisation of transportation networks.

The analysis of communication links between cells, regions or people can be used to observe the flow of information throughout a country. Such information makes it possible to identify socially connected groups. This has several marketing applications as well as providing insights into socially connected geographical regions. Applications focused on communication network optimisation may also use such data to help streamline service delivery.

In future chapters, techniques and algorithms which aim to solve some of the technical challenges associated with the aforementioned applications are developed. The specific focus is on the extraction of movement related behaviours that enable prediction of travel paths taken between points of interest, national mobility prediction, subscriber mobility routines and dynamic population estimation.

CHAPTER 4

Travel Path Discovery

Transportation surveys, which gather data on human mobility patterns and transport infrastructure utilisation, are widely used as inputs to strategic planning exercises by city [185], regional [186] and national [4] authorities with responsibility for the provision and maintenance of infrastructure and services for the general public. They are also a valuable data source for research into human mobility preferences [187] and societal trends [188]. Practically, however, transportation surveys are resource intensive and expensive to undertake as they typically involve either the deployment of dedicated monitoring hardware systems or manual data collection. Consequently, they tend to be performed infrequently, are of limited duration and only collect data for a small number of representative locations within the transportation network. Thus, while transport surveys are immensely valuable, they provide poor temporal and spatial resolution. This motivates the requirement for low cost and scalable alternatives. Exploiting the data collection capabilities of mobile phone networks is an attractive proposition in this regard [71, 147, 146, 149, 150, 189].

However, as the spatial accuracy of CDR positional estimates are restricted by network cell topology, it is often difficult to identify the particular route an individual travelled along between points of interest. Thus, in the absence of trip metadata, it is often quite difficult to attribute the flow of people between regions to particular routes, or distinguish modes of transport taken.

This chapter outlines novel research into the identification of mobile subscriber travel paths which attempts to address these issues. New similarity metrics are developed to quantify the resemblance of CDR trajectories and known travel paths between regions of interest. These metrics are compared to traditional trajectory comparison techniques, and are shown to outperform them when classifying the travel routes of trajectories within a test dataset. The test dataset, comprised of simulated journey trajectories, are generated using a novel agent based model which simulates CDR journey trajectories between points of interest. A novel methodology is then presented which identifies the routes taken by individuals as they travel between regions of interest. An overview of this procedure is depicted in Figure 4.1.

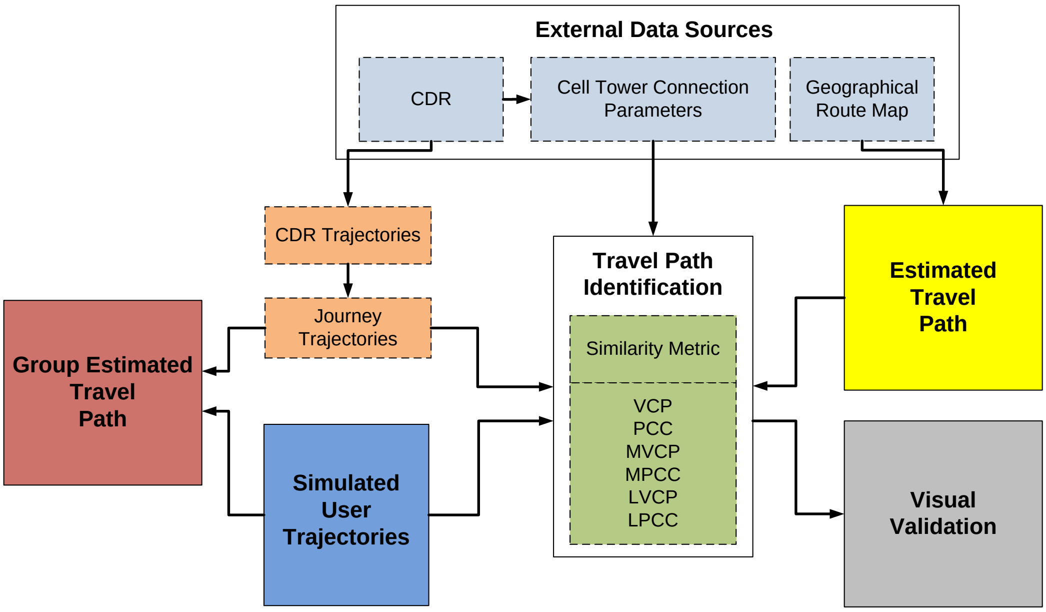


Figure 4.1: Methodology used to identify the routes taken by individuals as they travel between regions of interest.

Each stage of this procedure is detailed in the following sections of this chapter: Section 4.1 describes the formation of journey trajectories. Section 4.2 outlines a novel method for visualising the density of CDR journey trajectories, which is used during the visual validation of classification results. The details of the agent based model which simulates CDR journey trajectories between points of interest is given in Section 4.3. Section 4.4 then summarises each of the new similarity metrics proposed. A brief overview of traditional trajectory similarity measurement techniques is also given. Each of these techniques is then compared in Section 4.5 using a simulated CDR journey dataset. A novel procedure used to generate travel paths between regions of interest is then presented in Section 4.6. A classification algorithm which identifies the routes taken by subscribers as they travelled between Dublin city and Cork city is also introduced. Section 4.7 describes a novel procedure used to estimate travel paths belonging to a group of similar CDR trajectories. Finally, Section 4.8 concludes the chapter with a discussion outlining both the benefits and limitations of the proposed techniques for mobile travel path identification.

4.1 CDR Journey Trajectories

As previously discussed in Section 3.4, CDR trajectories are paths taken by mobile devices through a cell network as a function of time as observed from CDR. These trajectories may be segmented into CDR journey trajectories by identifying trajectory segments which correspond to trips between spatial regions of interest. A spatial region of interest, R, is a symbolic representation of an area in terms of cellular network coverage. Here, cells are deemed to belong to a region of interest if their site location falls within the spatial bounds of that location.

Trajectory path information is then sorted and preprocessed, as in Table 4.1, for future analysis, where m is the table placement index, u is the anonymised subscriber ID, i is the journey index, R_s is the starting region, R_e is the ending region, t is the time stamp of the activity, A is the recorded CDR activity, and C is an index to the serving cell tower. For clarity, the term activity and event are used interchangeably throughout. A journey trajectory (J) is defined as a single recorded path taken between two regions of interest, where each user may have several associated journey trajectories. An example of several extracted journey trajectories is depicted in Figure 4.2.

Ensuring anonymity of individual users whose trajectories are being tracked is important. Researchers have shown that the combination of trajectory data with external information sources can reveal the identity of previously unidentifiable users [80, 190, 191, 78, 79]. To address this issue, the index u is removed before analysis and journey identification is directly obtained from i. Such re-anonymisation reduces the likelihood of being able to identify any user associated with a particular journey, as the omission of a user reference breaks the linkage between blocks of travel information.

Table 4.1: Processed CDR structure used in the formulation of user CDR trajectories

m	u	i	R_s	R_e	t	A	C
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:

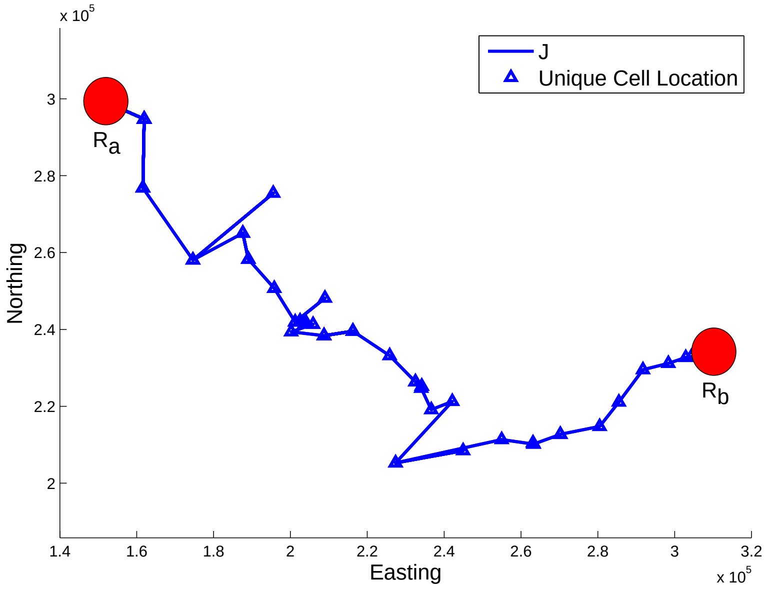


Figure 4.2: A CDR journey trajectory between two regions of interest, R_a and R_b.

4.2 Kernel Density Estimate (KDE)

To order to characterise the density of CDR trajectories, a suitable model is required which can accurately capture the likely travel path taken by a subscriber as they move from one area to the next. A number of mobility models have been developed for ad-hoc wireless networks [192, 193] and imperfect trajectory data [194]. Using accurate positional estimates, Demšar *et al* [195] outlined a suitable technique to estimate the kernel density of trajectories.

In the absence of an accurate model for CDR trajectory representation, plotting the density of the visited cell towers which make up a journey trajectory gives a good approximation of density as it captures the locations visited between the start and end locations of that journey. As discussed in Section 3.3, cell tower activities may be spatially distributed as a function of each cell tower's spatial parameters. Using a similar approach, a technique to estimate the kernel density of CDR trajectories is introduced.

The spatial smoothing function applied constructs an individual Gaussian bell for each unique cell tower present in a CDR journey trajectory, located at cell coverage area centroids

(C_x, C_y). The spreading factor of each Gaussian is governed by each cell tower radius, C_r, and spreads a weighting over a spatial lattice, $\varphi(x, y)$, such that the combined weight under an individual Gaussian sums to 1. The estimated kernel density is then obtained by combining each individual cell tower lattice. The spreading function used for a single cell tower is given by

$$\varphi(x, y) = \varrho \exp\left(-\frac{(x - C_x)^2}{2C_r^2} - \frac{(y - C_y)^2}{2C_r^2}\right), \tag{4.1}$$

where (x, y) are coordinates of points in the spatial lattice, C_r and (C_x,C_y) correspond to the cell radius and centroid location, and ϱ is a scaling weight which ensures that the combined weights in $\varphi(x, y)$ sum to 1. The combined weighted lattice, $\phi(x, y)$, is given by

$$\phi(x, y) = \sum_{C=1}^{N_J} \varphi_C(a, b) \tag{4.2}$$

where $\varphi_C(x, y)$ is the lattice for the cell tower indexed by C and N_J is the number of unique cell towers making up J. An illustrative example depicting a kernel density estimate for a single journey trajectory, J, shown in Figure 4.2 is given in Figure 4.3.

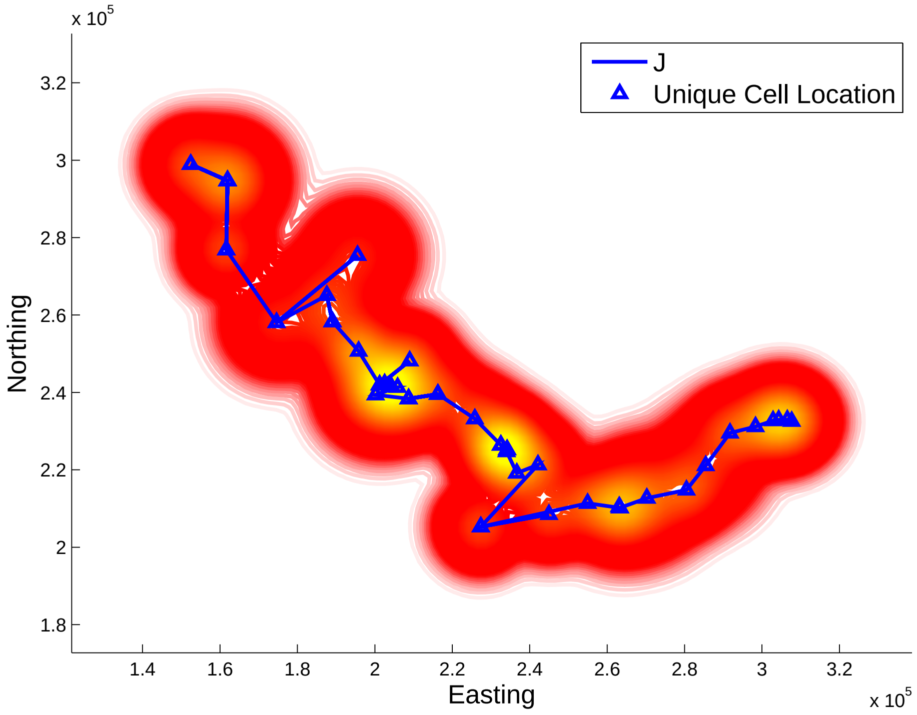


Figure 4.3: An illustrative example depicting a kernel density estimate for a single journey trajectory as shown in Figure 4.2.

4.3 Simulating CDR Journey Trajectories

This section outlines the development of an agent based model that combines mobile network cell tower connection probabilities, mobile phone activity distributions and geographical route map information, to model observed mobile subscriber activities along travel paths. The model outputs simulated journey CDR trajectories (J_i) which are constructed to contain sampling points along known travel paths between regions of interest such that they are, probabilistically speaking, representative of the actual subscriber trajectories observed through CDR.

4.3.1 Allocating Activity Locations

As previously discussed in Section 3.4, CDR trajectory sampling distributions are non uniform and are dictated by the activity profiles of individual subscribers. Within each trajectory, these distributions dictate activity locations, (A_x, A_y). Thus, a simulated journey trajectory needs to distribute activities along the selected travel path such that the bursty nature of device activity is maintained. As modern mobile devices may preform several activities simultaneously, the sampling distributions outlined in Section 3.4 are combined into one single distribution of τ. This distribution is displayed in Figure 4.4.

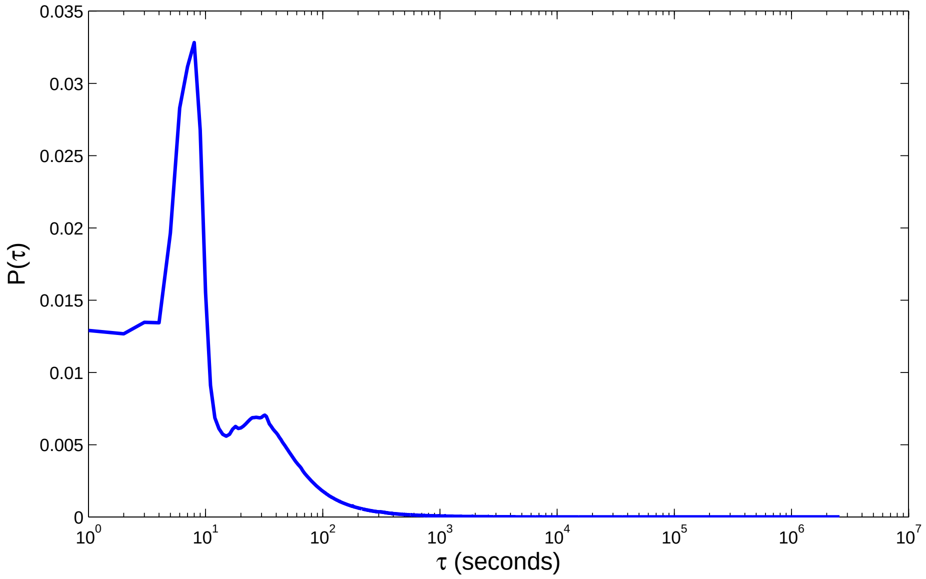


Figure 4.4: Probability density function (pdf) of time intervals between consecutive mobile phone activities

Given an initial start time of t_{st}, the temporal activities of users can be simulated by consecutively sampling the distribution of τ. However, there is a natural bias for samples to occur just after t_{st}. This feature is clearly evident from examining Figure 4.5. This figure illustrates the proportion of 1000 agents which where active during each temporal slot, given the time between each agent's activity is sampled from the distribution of τ. Thus it is clear that a settling time is needed which avoids the initial peak in agent activity. If not introduced, there is an assumption that every agent has performed an activity at t_{st}. Setting t_0 as the journey start time and t_{tot} as the total journey time, where $t_0 \gg t_{st}$, the temporal sequence of activities attributed to a journey may be selected from each agent's simulated temporal activity when,

$$t_0 \leq A_t \leq t_{tot} \tag{4.3}$$

where A_t is the estimated time of each activity, and is given by

$$A_{tk} = t_0 + \sum_{i=1}^{k} \tau_i \, . \tag{4.4}$$

Assuming a constant travel speed and total journey length L_{tot}, the relationship between distance travelled along the path and time passed is denoted by,

$$\frac{L_k}{L_{tot}} = \frac{A_{tk}}{t_{tot}} \, , \tag{4.5}$$

where A_{tk}/t_{tot} is the proportion of time passed since t_0 up until the activity indexed by k and L_k/L_{tot} is the corresponding proportion of distance travelled along the route. An example illustrating the selected locations of activities using this technique is given in Figure 4.6.

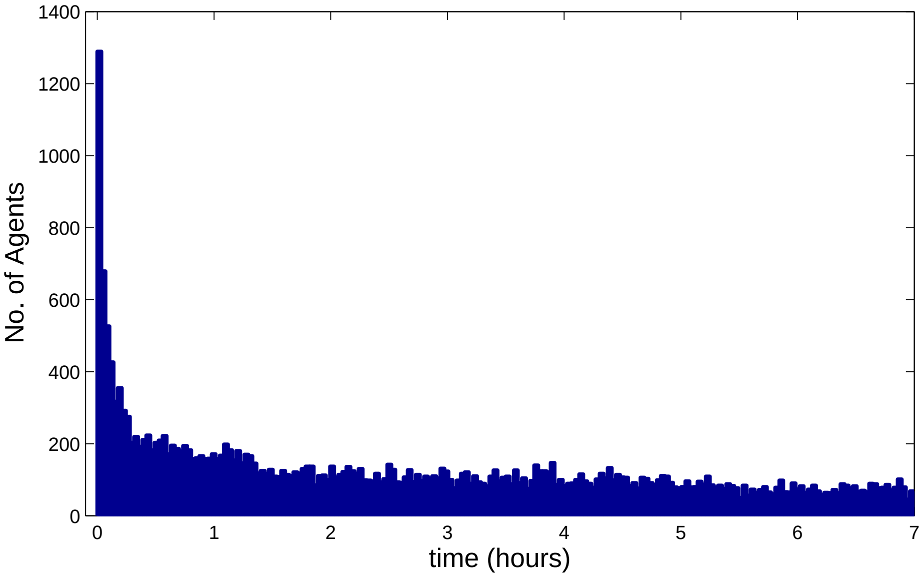


Figure 4.5: The number of agents active in each temporal slot given that the time between each activity is sampled from the distribution of τ, plotted in Figure 4.4. The total number of agents is 1000 and the temporal bin width is 2.1 minutes.

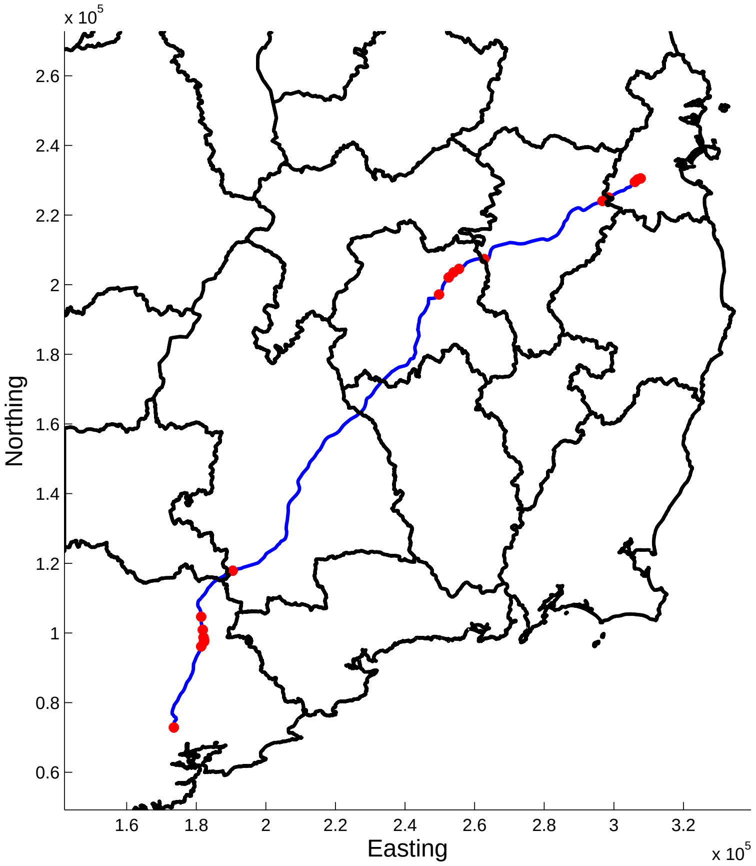


Figure 4.6: Estimated location of activities (A_x, A_y), indicated by ●, along a travel path of interest.

4.3.2 Cell Tower Selection

Once the location of each activity is estimated, the next step is to select the servicing cell towers. In real world situations, the selection of the cell which enables a call, SMS or data transfer to a mobile device is influenced by factors which include, among others, cell network topology, cell congestion and the perceived distance between the cell tower and device, which takes account of current channel characteristics and transmission power. To model this phenomenon, cell tower connection probabilities are derived from all relevant GPS-tracked mobile phone cell tower connection traces sourced from OpenCellID [196]. Cell tower connection probabilities allow for the deterministic localisation of a mobile device by isolating the most probable area the device was located while being serviced by that cell. This measure is then used to evaluate the likelihood of a cell servicing an activity which occurred at the location (A_x, A_y). By evaluating this measure for each cell in the network of interest, the selection of which cell services the activity may be determined stochastically.

The OpenCellID [196] database allows users to upload their GPS locations with their corresponding detected cell tower IDs via a client-side application installed on their mobile devices. Data from this database for the network under investigation amounted to just over 21,000 entries. These entries spanned the entire countryside, mostly along main roads and in cities, although a significant proportion were located in rural areas. The spatial distribution of activities is displayed in Figure 4.7.

To form cell tower connection probabilities, the distance between OpenCellID GPS recordings and the cell tower to which they are connected is tabulated. Each measurement may then be collated into a single distribution which describes the observed distances users were from cell towers while being serviced by them. If sufficient data were available, it would be possible to derive cell tower specific connection distributions from the GPS sourced information. In practice, the low number of samples and the biased sampling nature of the GPS recordings (e.g. many collected along roads but few for forests, etc.) means that individual connection distributions are not suitable for inference purposes. Instead, using the aforementioned distance measurements, it is possible to estimate a generic probability density function (pdf) which measures the likelihood of connecting to a cell as a function of the distance from the cell tower location. This distribution is displayed in Figure 4.8. Note the distribution of connection distance s is in general unique to each cell tower, reflecting the practical factors influencing their connection characteristics.

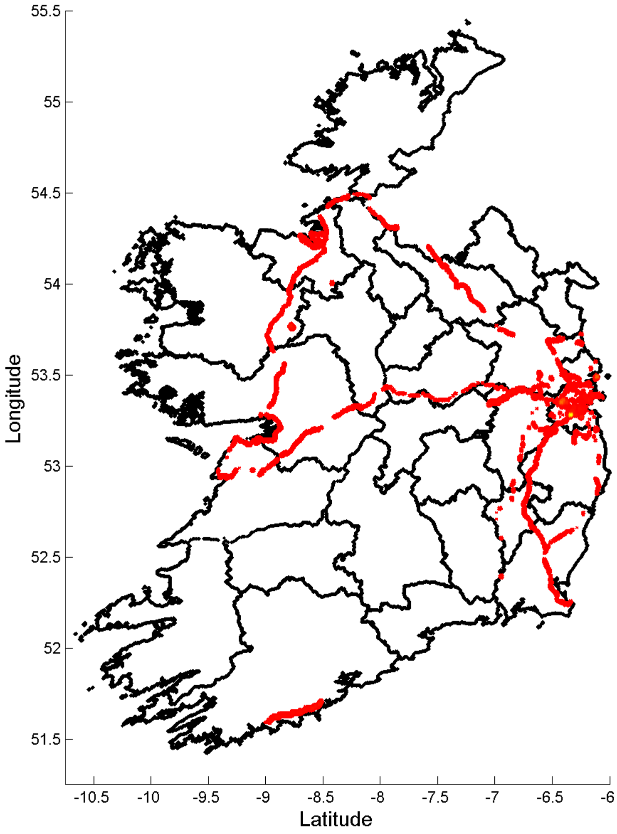


Figure 4.7: Distribution of GPS recordings from OpenCellID from devices active on the Meteor Network

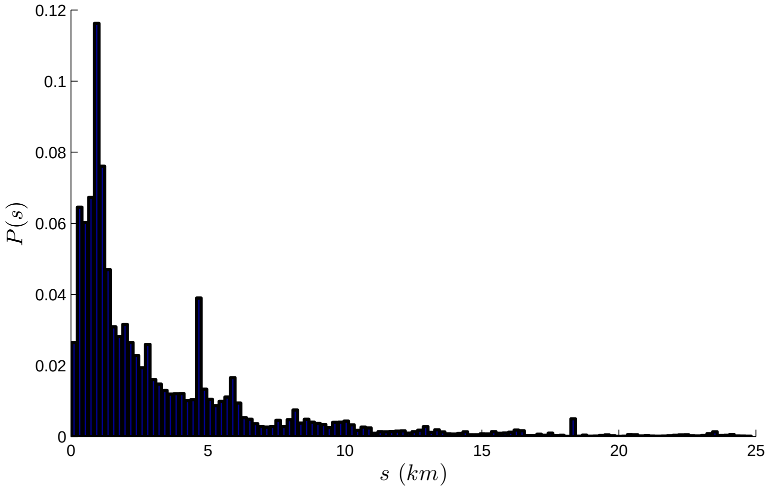


Figure 4.8: Distribution of the observed distances OpenCellID recordings were from servicing cell site locations.

A commonly observed phenomenon with GPS recordings is for repeated samples to cluster into dense regions once the device is stationary [197]. These dense regions are clearly evident within OpenCellID recording, as illustrated in Figure 4.9, and do introduce peaks within the derived pdf (i.e. the peak at the 5 km mark in Figure 4.8). As a result, it is necessary to remove such areas as they unduly bias particular distances. To isolate these regions, samples were grouped using DBSCAN [179], a density based clustering algorithm with noise. Each identified cluster was then replaced with a single point measurement located at its centroid. The aforementioned pdf is then reconstituted and is as depicted in Figure 4.10.

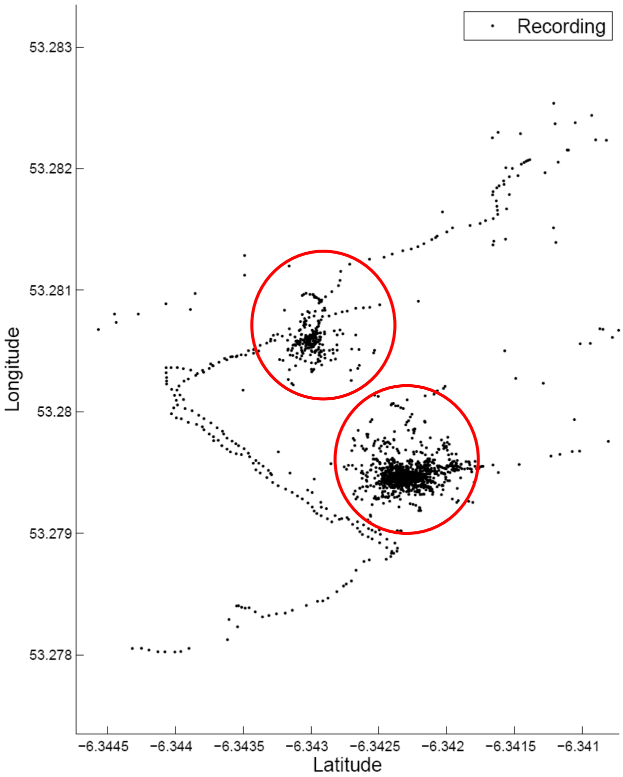


Figure 4.9: Irregularities (red circle) in OpenCellID recordings

A single distribution characterising the probability of connecting to a cell, which accounts for varying cell size, is constructed by scaling each OpenCellID recording by the theoretical radius of the connecting cell tower (C_r), as estimated from the Voronoi cell polygons. The resulting pdf, illustrated in Figure 4.11, measures the likelihood of connecting to a cell as a function of the normalised distance $\hat{s}$ from the cell tower location, where

$$\hat{s} = \frac{s}{C_r}. \tag{4.6}$$

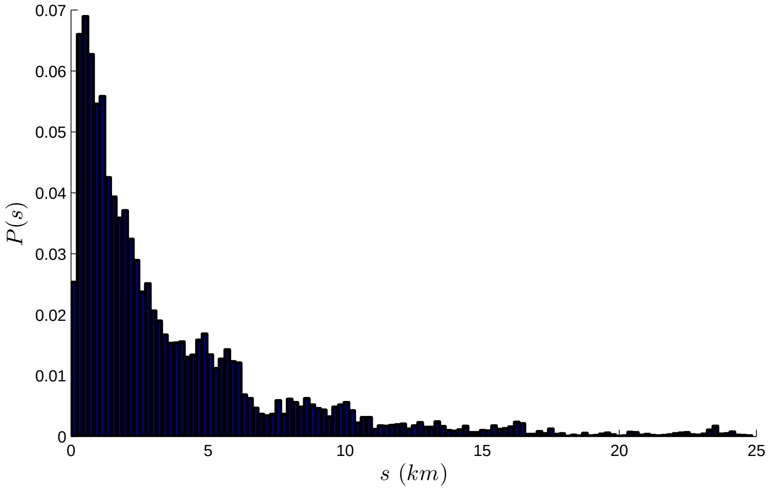


Figure 4.10: Distribution of the observed distances OpenCellID recordings were from servicing cell site locations with irregularities removed.

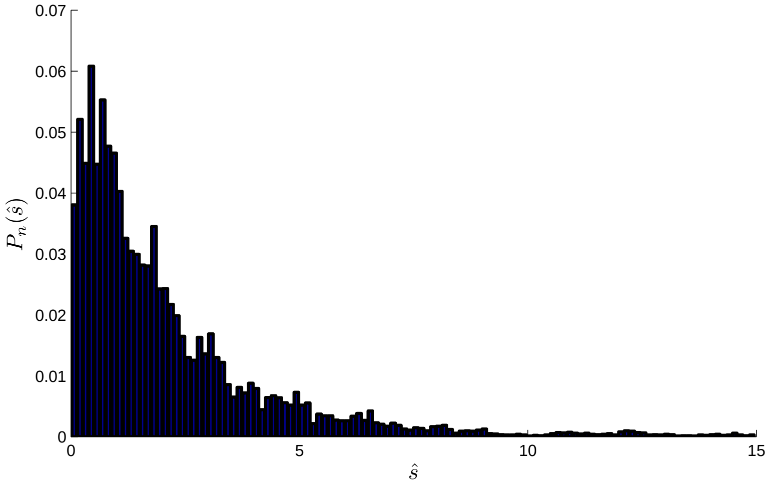


Figure 4.11: Distribution of likelihood of connecting to a cell as a function of the normalised distance $\hat{s}$ from the cell tower location.

Using this pdf, the probability of connecting to a cell while being at least a distance s from a cell with connection radius C_r is given by

$$P(\hat{s} > s/C_r) = P(\hat{s} > s^*) = \int_{s^*}^{\infty} P_n(\hat{s})ds \tag{4.7}$$

Given that connections can be assumed to occur on a 2D plane, the probability density function for connecting at a given radius ψ and a bearing θ to the cell tower C can be expressed as

$$P_C(\hat{\psi}, \theta) = \frac{P_n(\hat{\psi})}{2\pi\hat{\psi}}, \tag{4.8}$$

where $\hat{\psi} = \psi/C_r$ and ψ is the Euclidean distance from the activity location (A_x, A_y) and cell centroid (C_x, C_y). Angle θ is as depicted in Figure 4.12.

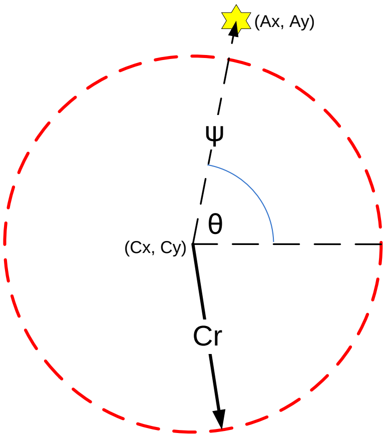


Figure 4.12: Illustration showing the calculation of distance ψ and angle θ.

The probability of an event occurring within a selected region or area, A_β, is given by $\int_{A_\beta} P_C(\hat{\psi}, \theta) dA_\beta$. This can be approximated by summing over an evenly distributed grid of discrete point measurements which fall within the enclosed region, that is

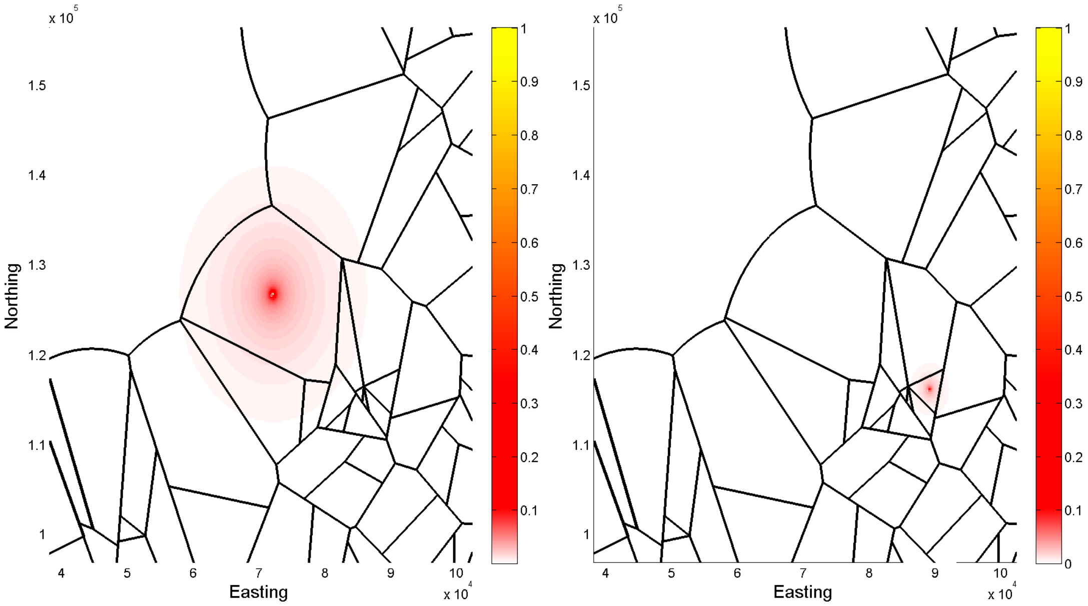
$$P(A_\beta) \approx \sum_{i=1}^{M_A} P_C(\hat{\psi}_i, \theta_i)\Delta_i \tag{4.9}$$

where Δ_i is the area of the i^{th} grid pixel, $(\hat{\psi}_i,\theta_i)$ are the polar coordinates of pixel centre and M_A is the number of points which fall within area A_β. For a uniform grid $\Delta_i = \Delta \; \forall \; i$ and hence

$$P(A_\beta) = \Delta \sum_{i=1}^{M_A} P_C(\hat{\psi}_i, \theta_i). \tag{4.10}$$

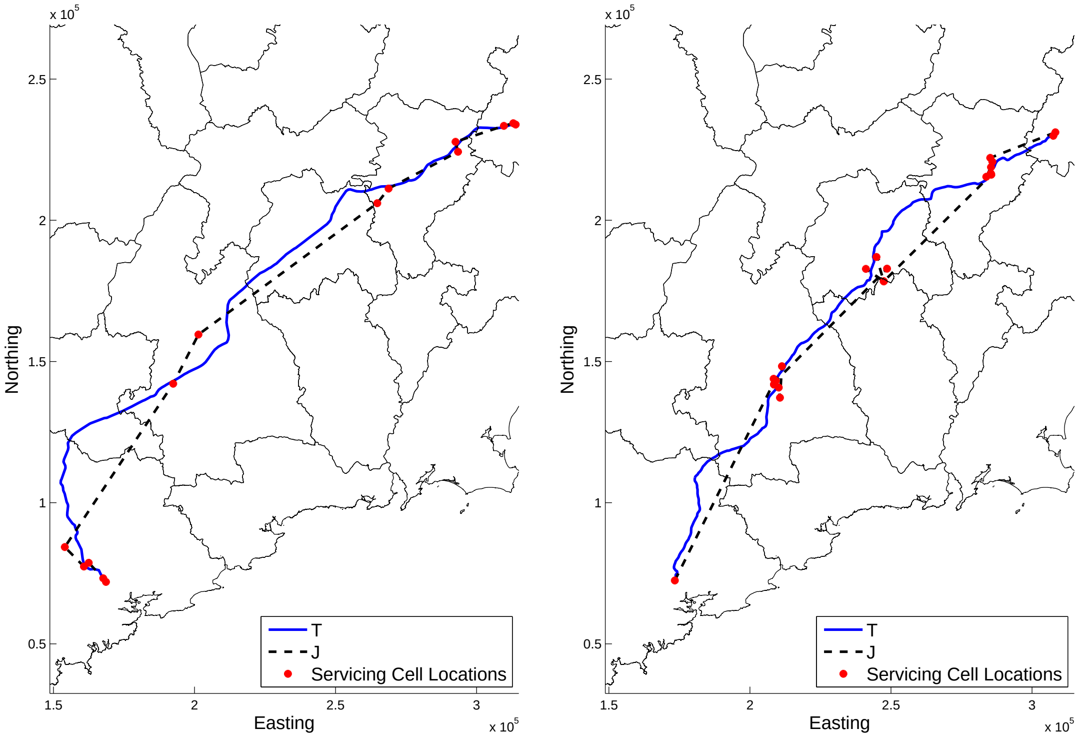
An example illustrating the spatial dispersion of connection probability from two cells of varying radius is depicted in Figure 4.13. To enable a direct comparison, each figure has been normalised to the same scale.

Given the position of an activity (A_x, A_y) along a travel path, as determined in Section 4.3.1, $P_C(\hat{\psi}_i, \theta)$ for each cell in the network under investigation is tabulated. The choice of which cell to select is then stochastically sampled from 50 of the top ranked cell towers. Fifty cells are selected as a compromise between spatial variance and realistic network behaviour. An example of two simulated CDR trajectories is depicted in Figure 4.14.



(a) (b)

Figure 4.13: The spatial dispersion of connection probability from selected cells: (a) Cell with a radius of 2 km; and (b) Cell with a radius of 10 km.



(a) (b)

Figure 4.14: Simulated CDR trajectories, J, travelling along travel paths of interest, T, between Dublin City and Cork City: (a) Path following rail line; and (b) Path following motorway.

4.4 Path Similarity Measurements

Path similarity is a quantitative measure that describes how similar two paths are. In this section, two novel similarity measures are introduced which measure the distances between CDR trajectories and travel paths. The first measurement technique involves virtual cell paths (VCP). This measurement is based on the proportion of events which occur at cells that are deemed to represent a route of interest. The second is based on probabilistic cell connectivity (PCC). PCC is a stochastic distance measurement which calculates the probability of activities within a journey trajectory being along any travel path. These similarity measurements are compared to commonly used trajectory similarity measures in Section 4.5. Alternative measurements used in this comparison are also detailed within this section. Additionally, novel enhancements to the similarity measurement Longest Common Subsequence (LCSS) are introduced, which enable LCSS to account for the spatial-heteroskedasticity which is present within CDR trajectories.

4.4.1 Virtual Cell Path

A virtual cell path is a collection of cells which represents a pathway through a mobile telephony network along which a user may travel while on any particular route. The virtual cell path distance between a route of interest, $T = [tp_1, tp_2, \dots, tp_{N_T}]$, and CDR journey trajectory, $J = [jp_1, jp_2, \dots, jp_{N_J}]$, is defined as

$$D_{VCP}(T, J) = \frac{n_j}{N_T} \, , \tag{4.11}$$

where n_j is the number of samples of the journey trajectory J which occur at cells contained within the VCP of travel route T and N_T is the length of T. In general, a VCP is comprised of cells whose coverage areas intersect with the travel path considered. Additional cells associated with a travel path, as manually identified from training data, may be added to improve the spatial accuracy of the resultant VCP. Once completed, each VCP consists of a list of cells whose spatial coverage either coincides with part of the travel path or serves as a connecting cell to users travelling along the path.

An example of a constructed VCP is depicted in Figure 4.15. Also depicted in Figure 4.15

are two selected CDR journey trajectories, J_1 and J_2 with length of 5 and 6 activities, respectively. Given that the number of activities occurring at cells within the VCP are 2 and 3, respectively, the corresponding VCP distances are $D_{VCP}(T, J_1) = 2/5$ and $D_{VCP}(T, J_2) = 3/6$.

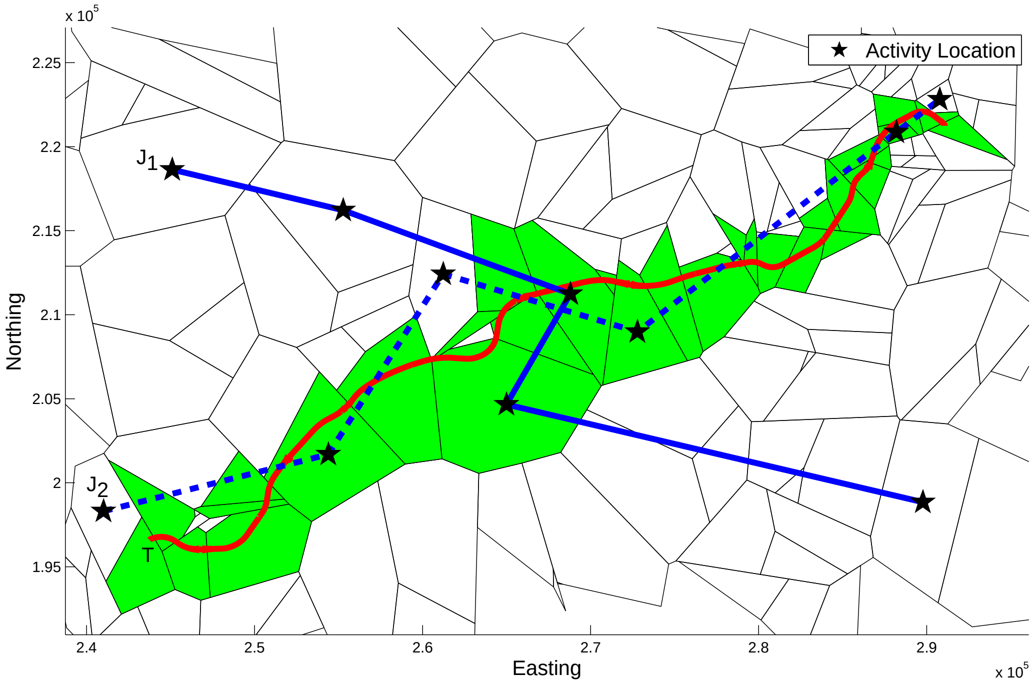


Figure 4.15: An example of a VCP and selected CDR journey trajectories, where the cells which construct the VCP for route T are indicated in green.

To account for the temporal sequencing of trajectories, the virtual cell path distance can be modified such that only temporally close trajectory points in J and T are compared. To add temporal data into T, we may apply a technique similar to that for distributed simulated activity locations along a route of interest, as detailed in Section 4.3.1. That is, each point in T is assigned a time proportional to the time it takes to reach that point while travelling along T given a constant speed and total travel time, where the total travel time is equal to the journey travel time of J. The modified virtual cell path distance, D_{MVCP}, between J and the modified T is given as

$$D_{MVCP}(T, J) = \frac{1}{N_j} \sum_{k=1}^{N_j} D_{VCP}(T^a, jp_k) \tag{4.12}$$

where $T^a = [tp_{k-\delta}, \ldots, tp_{k+\delta}]$ is a set of points in T temporally within δ of jp_k.

4.4.2 Probabilistic Cell Connectivity

The probabilistic cell connectivity distance, D_{PCC}, is a measurement of the probability that a CDR journey trajectory, J, can be attributed to a subscriber travelling along a particular route of interest, T. It is influenced by both the distance between T in relation to the connecting cells in J and the surrounding network topology of these cells. The distance measurement is given by

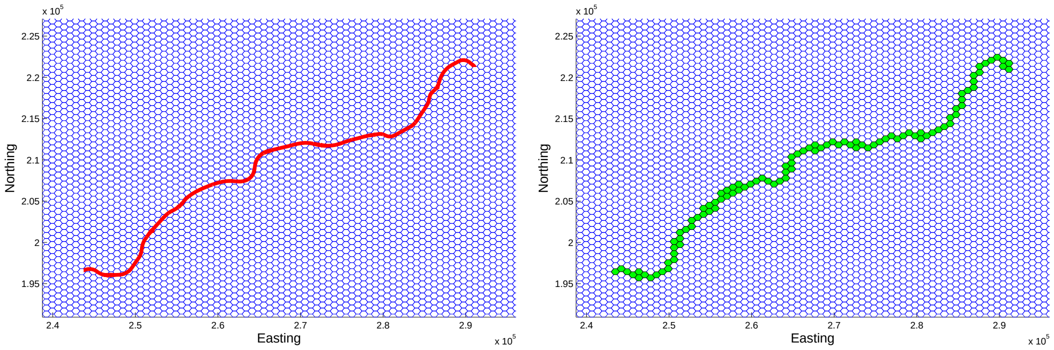
$$D_{PCC}(T, J) = \prod_{i=1}^{N_j} P(jp_i, T) \tag{4.13}$$

where N_j is the number of activities in a CDR journey trajectory, and $P(jp_i, T)$ is the probability that an activity at cell jp_i could be attributed to someone travelling along T.

If the trajectory of interest T was unevenly sampled, there might be undesirable biasing of $D_{PCC}(T, J)$ in certain spatial locations. To reduce the impact of such sampling bias, T is quantised using an evenly dispersed spatial lattice of hexagonal nodes. The quantisation method consists of selecting those nodes whose spatial coverage polygon intersects with the points in T. An illustrative example of the applied quantisation is depicted in Figure 4.16. Note the width of each tile in the lattice should be small enough to accurately capture the shape of T. The probability of a single activity at cell C corresponding to a travel path of interest T is then estimated as

$$P(C, T) = \sum_{j=1}^{N_q} P_C(\hat{\psi}_j, \theta_j) \tag{4.14}$$

where N_q is the number of quantised sample points which form T, and $P_C(\hat{\psi}_j, \theta_j)$ is calculated for each point as detailed in equation 4.8.



(a)

(b)

Figure 4.16: Quantisation of a trajectory of interest, T, onto a spatial lattice of hexagonal nodes.

By modifying $P(C,T)$ to account for the temporal sequencing of trajectory samples, a modified version of probabilistic cell connectivity distance (D_{MPCC}) can be derived. Applying the same temporal extension to T as discussed in Section 4.4.1, the modification to $P(C,T)$ is given as

$$P(C,T^a) = \sum_{j=1}^{n_j} P_C(\hat{\psi}_j, \theta_j)\,. \tag{4.15}$$

4.4.3 Hausdorff and Modified Hausdorff Distances

The Hausdorff distance refers to the greatest possible distance from any point in a trajectory to the closest point in another [198]. This measurement is defined primarily for unequal length data. However, it does not take account of the ordering of points, thus it may incorrectly match dissimilar trajectories (i.e. driving different directions on the same road) [20]. The Hausdorff distance between a route of interest, T and CDR journey trajectory, J is defined as

$$D_H(T,J) = \max\left(D_h(T,J), D_h(J,T)\right). \tag{4.16}$$

The distance $D_h(T,J)$ is given as

$$D_h(T,J) = \max_k \left(\min_l d_E(tp_k, jp_l)\right) \quad \forall k,l \tag{4.17}$$

where $d_E(tp_k, jp_l)$ is the Euclidean distance between a point in T, tp_k, and a point in J, jp_l.

The modified Hausdorff distance [199] was introduced to account for temporal ordering within trajectories. It is defined for T and J as

$$D_{MH}(T,J) = \max\left(g_d(T,J), g_d(J,T)\right) \tag{4.18}$$

where $g_d(T,J)$ is the given by

$$g_d(T,J) = \frac{1}{N_T}\sum_{tp\in T}\min_{jp\in J}\left(d_E(tp,jp)\right) \tag{4.19}$$

4.4.4 Dynamic Time Wrapping

Dynamic Time Wrapping (DTW) measures the distance between two trajectories of unequal length by finding a time warping that minimises the total distance between matching points [200] [201]. The DTW distance between a route of interest, T, and CDR journey trajectory, J, is defined as

$$D_{DTW}(T,J) = \frac{(d_{DTW}(T,J) + d_{DTW}(J,T))}{2} . \tag{4.20}$$

The distance $d_{DTW}(T,J)$ is given by

$$d_{DTW}(T,J) = \frac{1}{K}\sum_{k=1}^{K} d_E(\phi_{tp,k}, \phi_{jp,k})\, m_k / M_\phi \tag{4.21}$$

where ϕ_{tp} and ϕ_{jp} are the time warping functions that minimise the distance between aligned points, m_k is a path weighting coefficient, and M_ϕ is a path normalisation factor. The warping path ϕ can be efficiently found using dynamic programming.

4.4.5 Longest Common Subsequence

Similar to DTW, Longest Common Subsequence (LCSS) is a trajectory similarity measurement which aligns trajectories of unequal length [202]. However, it is more robust to noise and outliers than DTW because not all points need to be matched. Instead of one-to-one mapping between points, a point which does not have a good match may be ignored to prevent unfair biasing [201]. The LCSS distance between a route of interest, T, and CDR journey trajectory, J, is given by [201]

$$D_{LCSS}(T,J) = 1 - \frac{LCSS(T,J)}{\min(a,b)} \tag{4.22}$$

where $LCSS(T,J)$ specifies the number of matching points between two trajectories and is calculated as

$$LCSS(T,J) = \begin{cases} 0 & a = 0 \mid b = 0 \\ 1 + LCSS(T^{a-1}, J^{b-1}) & d_E(tp_a, tp_b) < \epsilon \;\&\; |a-b| < \delta \\ \max(LCSS(T^{a-1}, J^{b}), LCSS(T^{a}, J^{b-1})) & \text{otherwise} \end{cases} \tag{4.23}$$

where $T^a = \{tp_1, tp_2, \ldots, tp_a\}$ denotes all the flow vectors in trajectory T up to time a, and both ϵ and δ are distance and time thresholds, respectively. Like DTW, LCSS can be efficiently computed using dynamic programming.

4.4.6 Modified LCSS

Because $LCSS(T, J)$ uses Euclidean distance to evaluate whether or not a singular point in J is within ϵ of T, it is unable to account for CDR trajectory spatial-heteroskedasticity. This motivates the use of D_{VCP} or D_{PCC} as an enhancement for $LCSS(T, J)$. By replacing $d_E(tp_a, tp_b)$ with $D_{VCP}(tp_b, tp_a)$ or $D_{PCC}(tp_b, tp_a)$, $LSCC(T, J)$ may be modified to account for the spatial variance present in CDR trajectories.

$LCSS(T, J)$ with D_{VCP} for a route of interest, $T = \{tp_1, tp_2, \ldots, tp_{N_J}\}$, and CDR journey trajectory, $J = \{j_1, j_2, \ldots, j_M\}$, is given by

$$LVCP(T, J) = \begin{cases} 0 & a = 0 \mid b = 0 \\ 1 + LCSS(T^{a-1}, J^{b-1}) & D_{VCP}(t_a, t_b) > \epsilon \ \& \ |a - b| < \delta \\ \max(LCSS(T^{a-1}, J^b), LCSS(T^a, J^{b-1})) & \text{otherwise} \end{cases} \tag{4.24}$$

Likewise the modification $LCSS(T, J)$ with D_{PCC} is given by

$$LPCC(T, J) = \begin{cases} 0 & a = 0 \mid b = 0 \\ 1 + LCSS(T^{a-1}, J^{b-1}) & D_{PCC}(tp_a, tp_b) > \epsilon \ \& \ |a - b| < \delta \ . \\ \max(LCSS(T^{a-1}, J^b), LCSS(T^a, J^{b-1})) & \text{otherwise} \end{cases} \tag{4.25}$$

4.5 Comparative Study

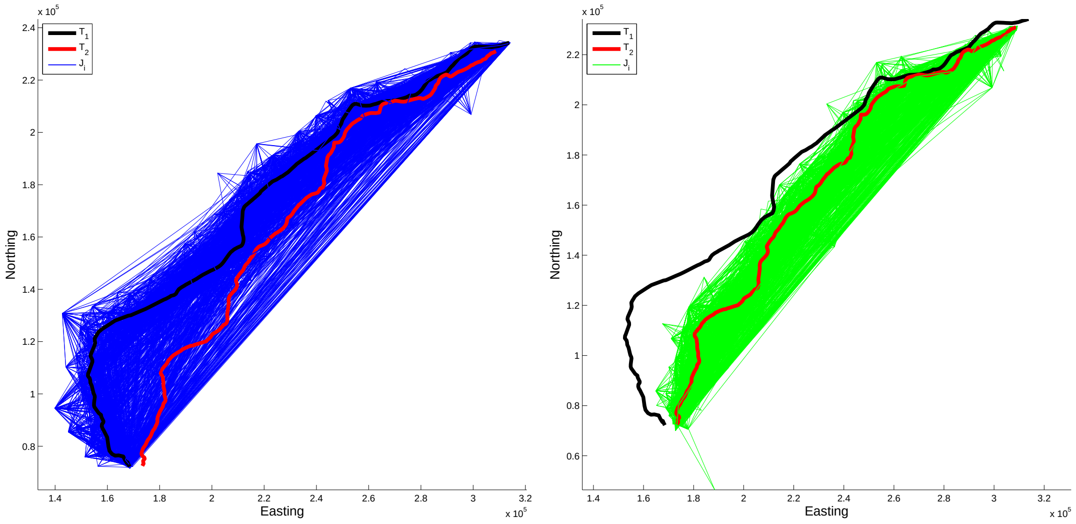
To evaluate the effectiveness of the D_{VCP} and D_{PCC} distance measurements proposed, a comparison is made among each technique and standardised methods which have been shown to be effective trajectory similarity measures. The trajectory similarity measurements used in the study are outlined in Table 4.2. Note, we restrict ourselves in this study to only spatial similarity as opposed to temporal or semantic feature similarity. This is a common practise as it results in a natural interpretation of spatial proximity [201]. The study is also restricted to a select group of suitable similarity measurements. For a more complete list of trajectory similarity measurements see Morris *et al.* [20, 201], Dodge *et al.* [203] and Zhang *et al.* [204].

The dataset used for the comparison consisted of 2,000 simulated CDR journey trajectories, split equally between two travel paths. The travel paths and corresponding simulated CDR journey trajectories are depicted in Figure 4.17. Using D_{VCP}, D_{PCC} and each of the similarity measurements as outlined in Table 4.2, the distance between each J and T was tabulated. Each

J was then assigned to the closest travel path. The results are presented in Table 4.3, and reflect the percentage accuracy of each distance measurement for the given dataset.

Table 4.2: Trajectory similarity measurement techniques.

Technique	**Reference**
Hausdorff Distance	Lou [198]
Modified Hausdorff Distance	Dubuisson and Jain [199]
Dynamic Time Wrapping	Keogh *et al.* [200]
Longest Common Subsequence	Hirschberg [202]



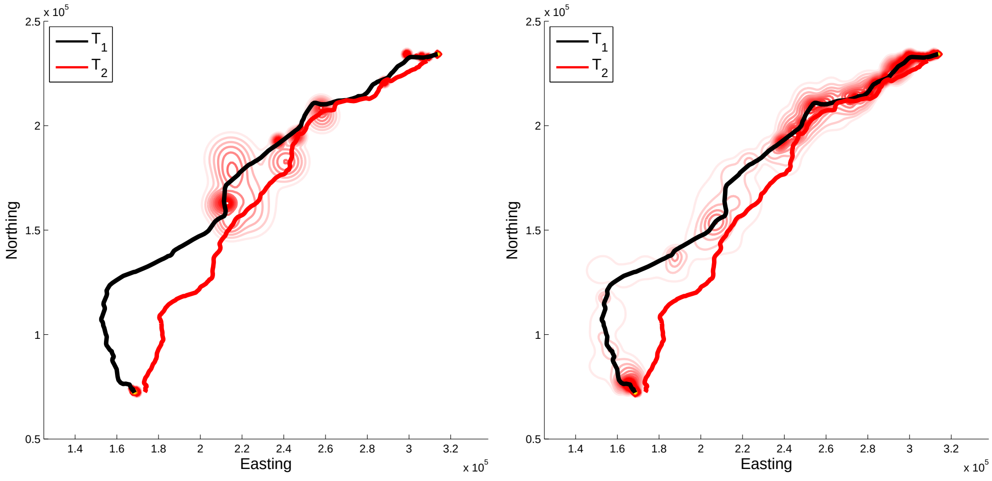
(a) (b)

Figure 4.17: Trajectory dataset used to compare the similarity metrics. The simulated trajectories (J_i) along travel paths T_1 and T_2 are depicted in (a) and (b), respectively.

Table 4.3: Accuracy of closed travel path assignment using different similarity measurements.

Technique	% T_1	% T_2	% $Total$
VCP	98.60	100	99.30
PCC	100	100	100
Hausdorff	45.40	97.80	71.60
Modified Hausdorff	93.80	75.10	84.45
Dynamic Time Wrapping	11.60	80.20	45.90
Longest Common Subsequence	86.10	99.40	92.75

The results show that D_{VCP}, D_{PCC} and D_{LCSS} are effective metrics for spatial proximity between paths of interest and simulated CDR trajectories. Note, the distance parameter ϵ used in D_{LCSS} has been optimised to achieve the highest classification accuracy for this dataset. Otherwise, no training was applied. D_{PCC} demonstrates that it is more effective at inferring subscriber travel paths compared to other techniques tested. Both D_{VCP} and D_{LCSS} tend to struggle when paths of interest are situated close together and occupy some common cell coverage areas. This is evident from viewing the spatial distribution of simulated activities from misclassified journeys as given in Figure 4.18.



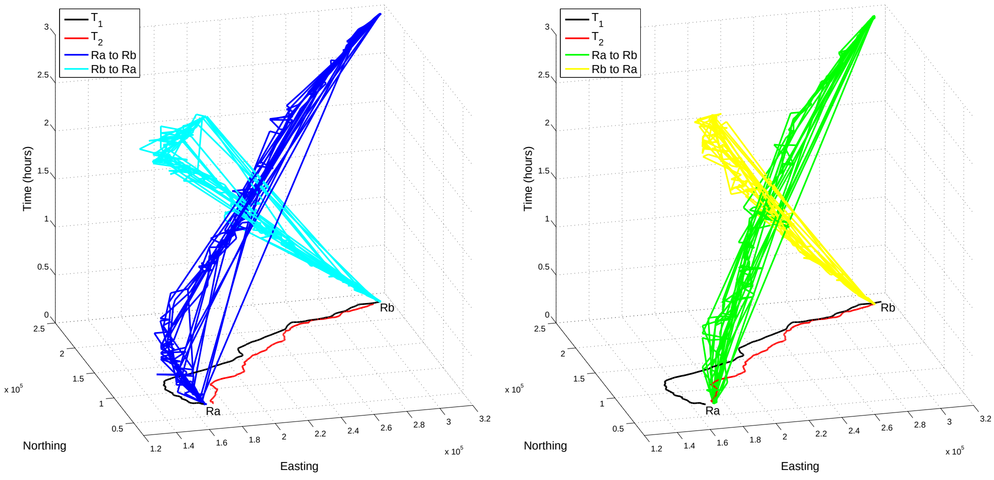
(a)

(b)

Figure 4.18: Kernel density estimate of simulated activities from misclassified journeys for (a) VCP; and (b) LCSS.

Without a temporal component, D_{VCP} and D_{LCSS} apply similar distance metrics, however, D_{VCP} has the added feature of accounting for surrounding cell topology which enables it to account for CDR trajectory spatial-heteroskedasticity. As previously discussed in Section 3.4, CDR trajectory spatial-heteroskedasticity refers to the variance in uncertainty of positional estimates, which, is a function of the physical topology of the mobile network. The accuracy of D_{PCC} reflects its ability to account for both cell topology and network connection mechanics in a spatial context. Using simulated CDR journey trajectories with both spatial and temporal features, a performance comparison of D_{MVCP}, D_{MPCC}, D_{LCSS}, D_{LVCP} and D_{LPCC} was performed. A sub-sample of the journey trajectories used in the comparison is depicted in Figure 4.19. The distance between the simulated CDR journey trajectories and each travel path was then tabulated in the direction of $R_a \rightarrow R_b$ and $R_b \rightarrow R_a$, with each trajectory being assigned to the closest travel path. If two or more travel paths were equidistant the CDR trajectory was marked as indistinguishable (Ind).

The percentage accuracy of each distance measurement for the given dataset are presented in Table 4.4. The results show that both D_{MVCP} and D_{MPCC} outperformed D_{LCSS}, D_{LVCP} and D_{LPCC} when determining the correct travel path taken. By incorporating D_{PCC} or D_{VCP} into D_{LCSS} improved accuracy with respect to CDR trajectory distance calculations can be observed. The high number of indistinguishable trajectories recorded with D_{LCSS} is due to the combination of D_{LCSS}'s inability to account for CDR trajectory spatial-heteroskedasticity, and the omission of points which fall outside temporal and spatial tolerances in the calculation of $D_{LCSS}(T, J)$. The omission of such points is a well documented feature of LCSS which enables it to account for noise within trajectories. As a result, both D_{LVCP} and D_{LPCC} will account for instances where noise is introduced to a J due to outdated cell tower locations. Therefore, there is a tradeoff between accuracy and noise tolerance when deciding which similarity metric to apply.



(a)

(b)

Figure 4.19: Sample of simulated CDR journey trajectories between regions R_a and R_b along: (a) T_1 and (b) T_2.

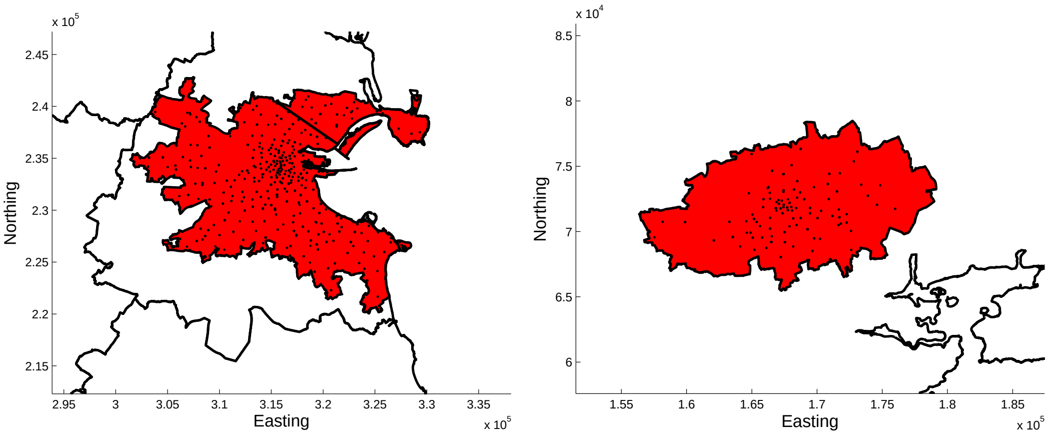
Table 4.4: Spatio-temporal trajectory similarity measurement results

Technique	% $\mathbf{T}_{\mathbf{1},R_a \to R_b}$	% $\mathbf{T}_{\mathbf{2},R_a \to R_b}$	% $\mathbf{T}_{\mathbf{1},R_a \to R_b}$	% $\mathbf{T}_{\mathbf{2},R_a \to R_b}$	**% Ind**	**% Er**
MVCP	98.80	99.00	100	99.80	0.60	0.00
MPCC	99.20	99.40	100	100	0.35	0.00
LCSS	86.20	66.60	66.80	76.25	22.95	0.80
LVCP	98.20	85.60	100	88.60	5.90	1.00
LPCC	100	77.20	98.40	66.80	14.20	0.20

4.6 Estimating Travel Paths

The ability to observe the travel paths taken by individuals as they migrate between regions of interest allows planning authorities to identify the particular routes which serve as gateways between each region. This information can be valuable when deciding on future infrastructural requirements, as planners can observe the number of individuals and the routes they take while travelling between towns or cities at a much higher temporal resolution compared to that currently possible with traditional travel survey methods. To demonstrate this an investigation was carried out which aimed to identify the travel paths taken by individuals who travelled between the Republic of Ireland's two largest cities, Dublin City and Cork City from 03-01-2011 to 10-01-2011, as observed through CDR.

For this purpose CDR journey trajectories were extracted for people who travelled between the two cities over the study period. The regions defining each cities boundary is depicted in Figure 4.20. The number of identified journeys over this time period was approximately 9500. Each CDR journey trajectory is displayed in Figure 4.21, with the distribution of journey times given in Figure 4.22. As can be clearly observed, individuals travelled along several different paths, with many travelling indirectly between each city. To identify the journeys taken directly between each city along known major transportation links (rail and motorway), journeys were classified based on their similarity to each transportation link trajectory. The transportation links are depicted in Figure 4.23.



(a) Dublin City

(b) Cork City

Figure 4.20: Cells located within city boundaries of (a) Dublin City; and (b) Cork City. Cell site locations are indicated by black dots.

For this purpose, we applied D_{PCC} as a similarity measure between each journey trajectory and transportation link and classified the journey type of each trajectory as follows:

$$T = \begin{cases} Rail & \text{if } D_{PCC}(T_{rail}, J) \; > \; D_{PCC}(T_{road}, J) + \epsilon \\ Road & \text{if } D_{PCC}(T_{road}, J) \; > \; D_{PCC}(T_{rail}, J) + \epsilon \\ unknown & \text{if } D_{PCC}(T_{road}, J) \; = \; D_{PCC}(T_{rail}, J) \; = \; 0 \\ Indistinguishable & \text{otherwise} \end{cases} \tag{4.26}$$

The results of this classification are given in Table 4.5. A kernel density of the journeys which were identified as travelling along each transportation link is given in Figure 4.24. The corresponding travel times are depicted in Figure 4.25.

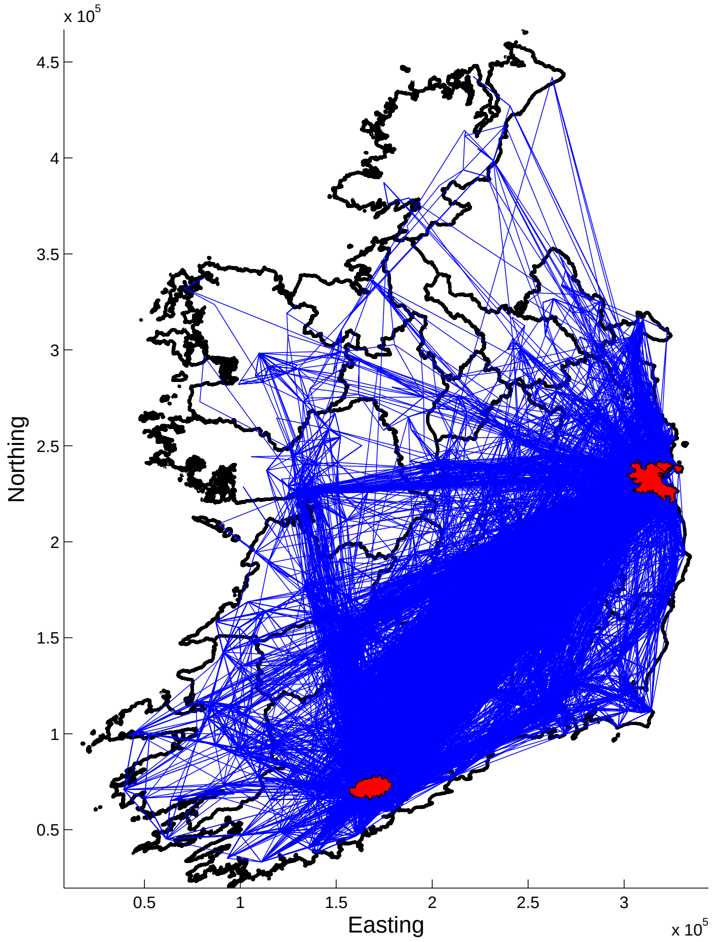


Figure 4.21: Extracted journey trajectories between Dublin City and Cork City from 03-01-2011 to 10-01-2011 where city boundaries are indicated by red polygons.

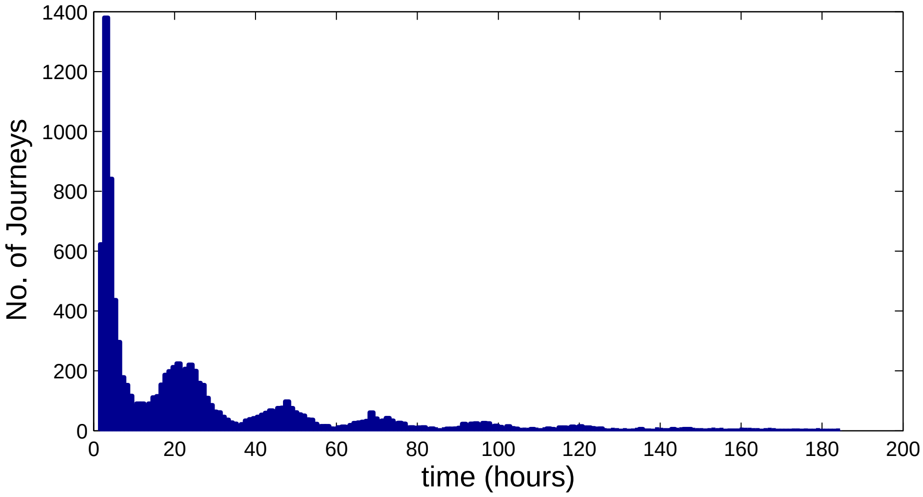


Figure 4.22: Travel time distribution of extracted journey trajectories between Dublin City and Cork City. Note a temporal bin width of 1 hour was used.

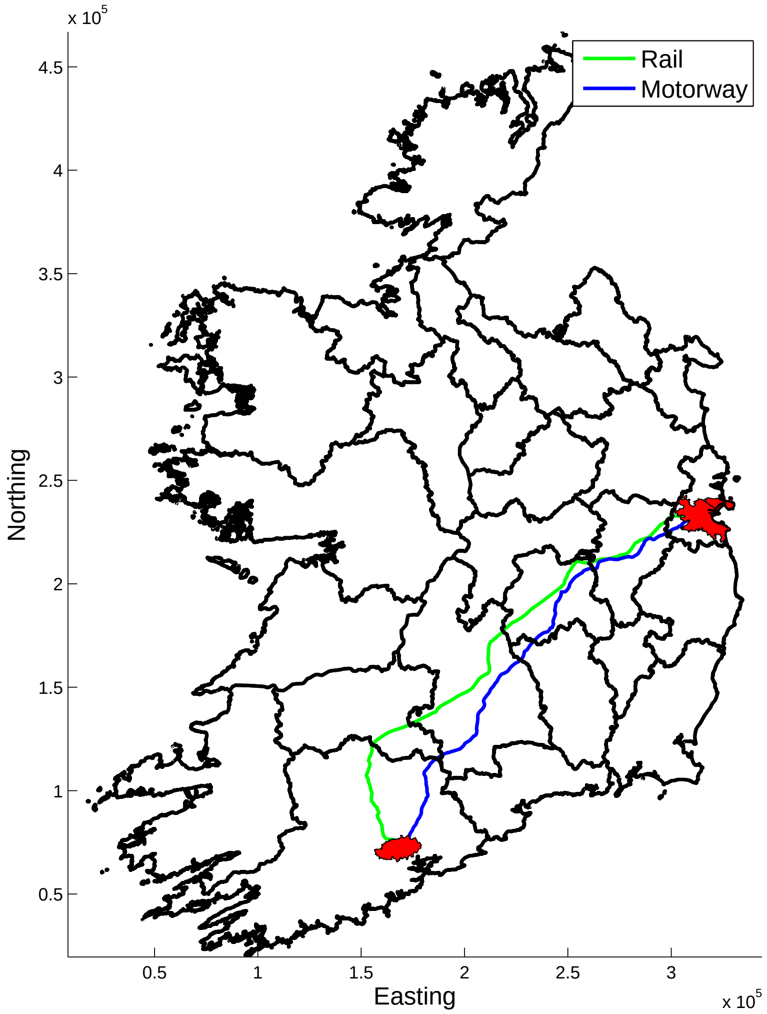


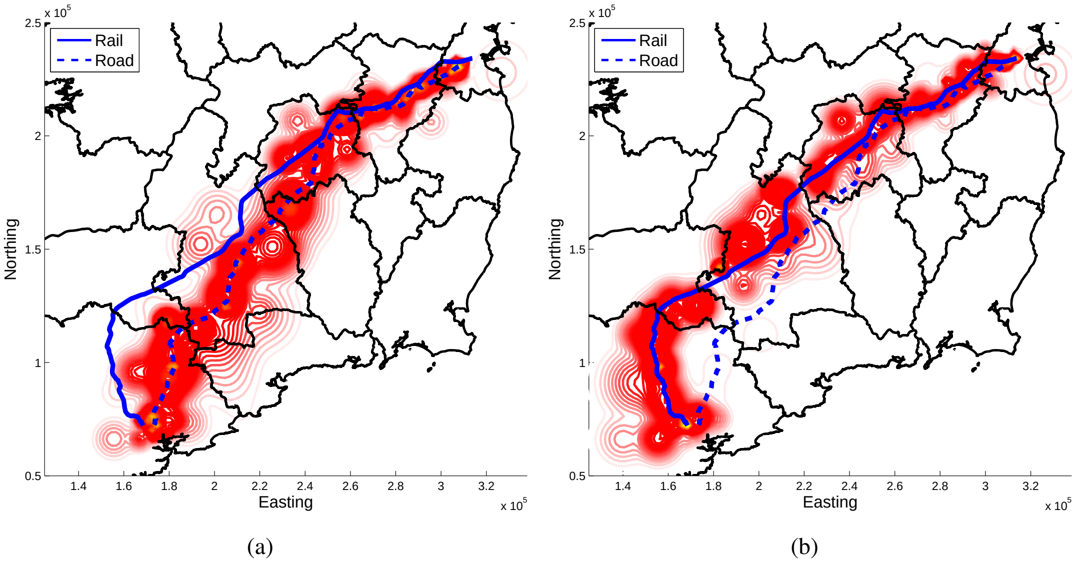
Figure 4.23: Major transportation links between Dublin City and Cork City.

However, a significant number of journeys were not classified as travelling along either direct route. To identify indirect paths, a novel methodology was developed which extracts alternative routes using geographical route map information from OSI [174]. This data consisted of unlabelled road and rail point data corresponding to the locations of road and rail tracks through the Republic of Ireland, as illustrated in Figure 4.26.

The first step involves quantising the point data. The method of quantisation is the same as that used by PCC, as outlined in Section 4.4.2. Like PCC, a hexagonal lattice is used as the underlying structure. This is chosen to maintain a constant spatial distance relationship between neighbouring nodes. An example lattice indicating nodes which contain road point data is shown in Figure 4.27.

Table 4.5: Classification of travel path taken by 9490 journey trajectories between Dublin City and Cork City.

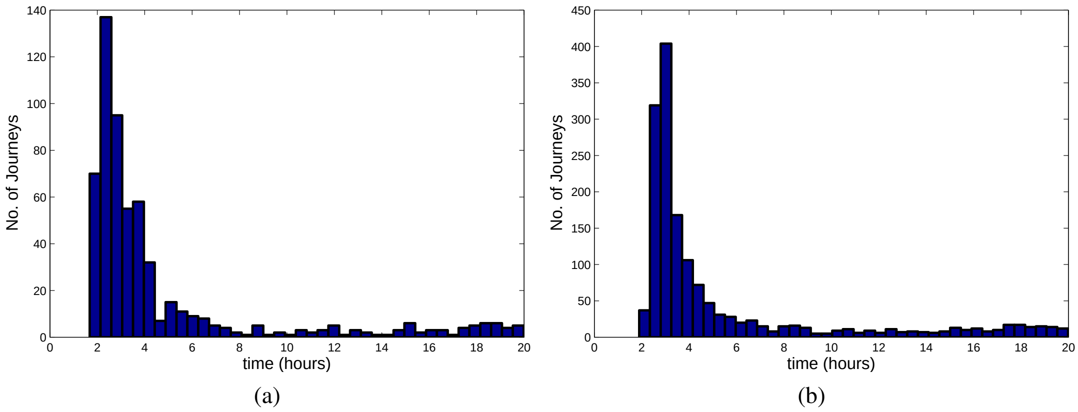
Travel Path	**Number**
Unknown	8475
Rail	732
Road	283
Indistinguishable	5



(a)

(b)

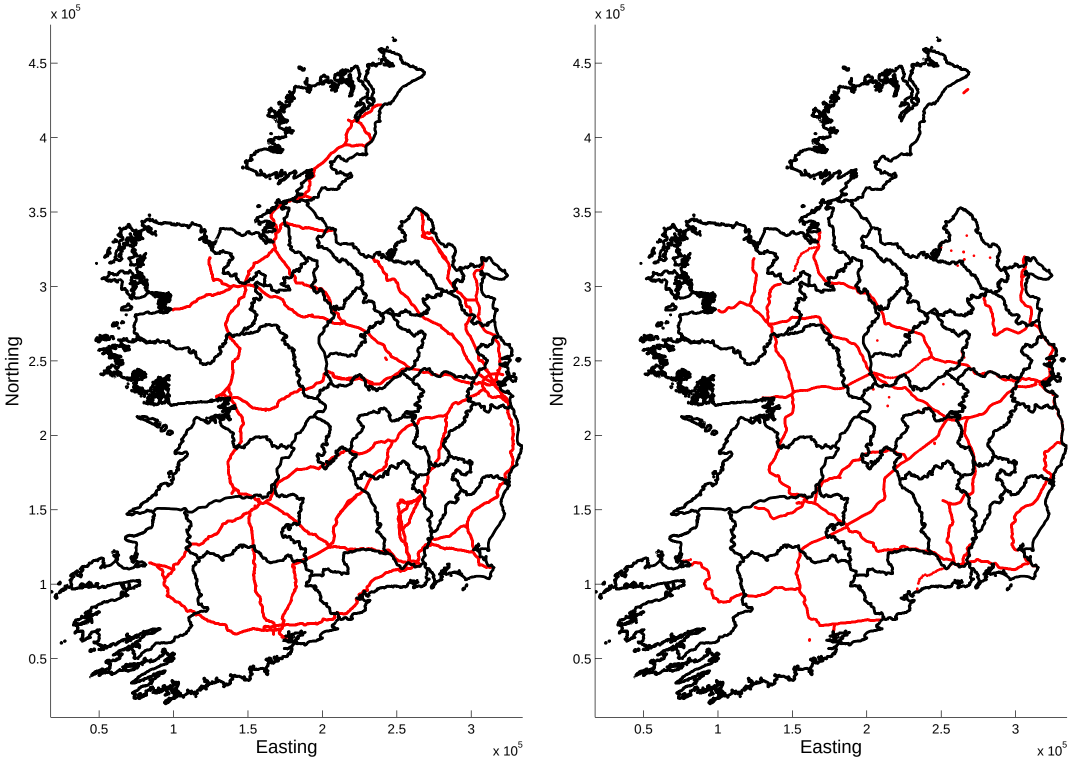
Figure 4.24: Kernel density estimate of journey trajectories identified as travelling along (a) road; and (b) rail travel paths.



(a)

(b)

Figure 4.25: Travel times of journey trajectory identified as travelling along (a) road; and (b) rail travel paths.



(a) Road

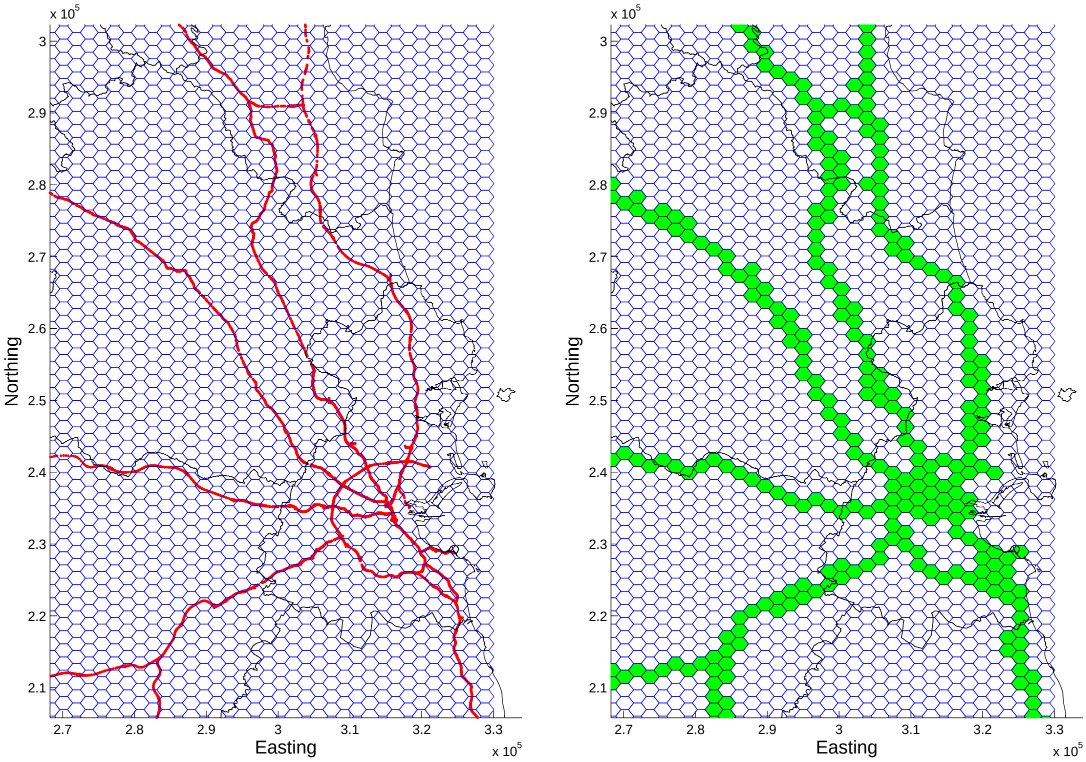
(b) Rail

Figure 4.26: Unlabelled points which correspond to the locations of (a) road; and (b) rail tracks.

The next step is to quantify the cost associated with moving between each node. This is achieved using a travel cost matrix given by, Γ,

$$\Gamma = \begin{pmatrix} \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,N_L} \\ \gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,N_L} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{N_L,1} & \gamma_{N_L,2} & \cdots & \gamma_{N_L N_L} \end{pmatrix} \tag{4.27}$$

where γ_{ij} is the cost associated with moving from node $i \rightarrow j$, and N_L is the number of nodes in a lattice. $\gamma_{ij} = 1$ for neighbouring nodes, and $\gamma_{ij} = \infty$ otherwise. This is done to restrict movement, and ensure only a transition to an adjacent neighbouring node is allowed in any one step, i.e. no jumping is allowed. Each row in Γ is then scaled by a Boolean vector σ, where $\sigma(i) = 1$ if node i contains route point data or $\sigma(i) = \text{inf}$ otherwise. The effect of this scaling is to further restrict movement, such that only a transition between two neighbouring nodes is permitted if both contain route point data.

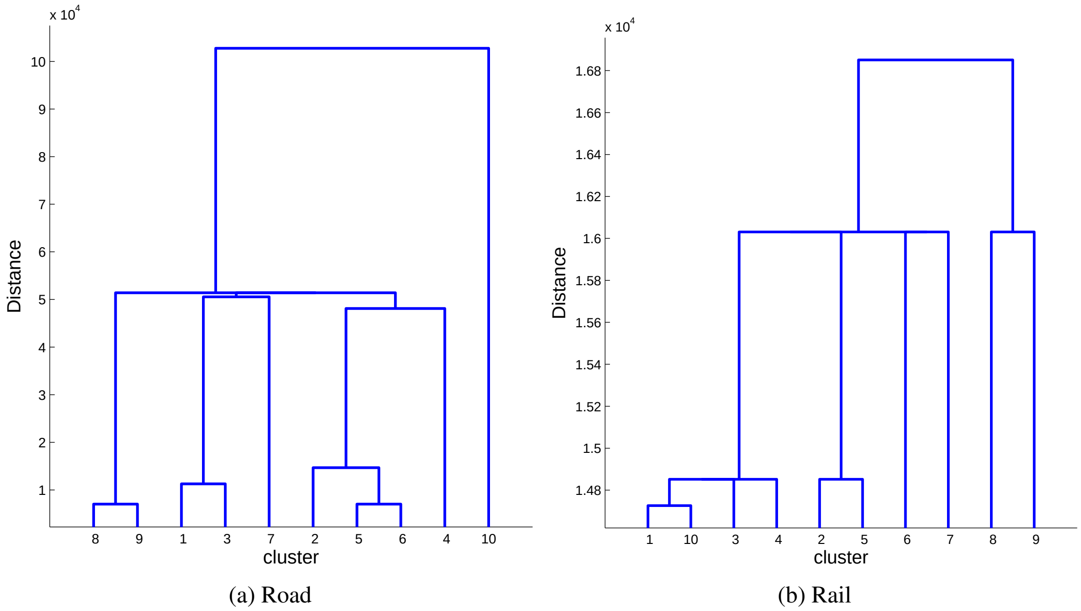


(a)

(b)

Figure 4.27: A sample of a hexagonal lattice which covers road point data; (a) road point data indicated in red; and (b) nodes containing road point data indicated in green. Note each lattice node width is 2.5 km.

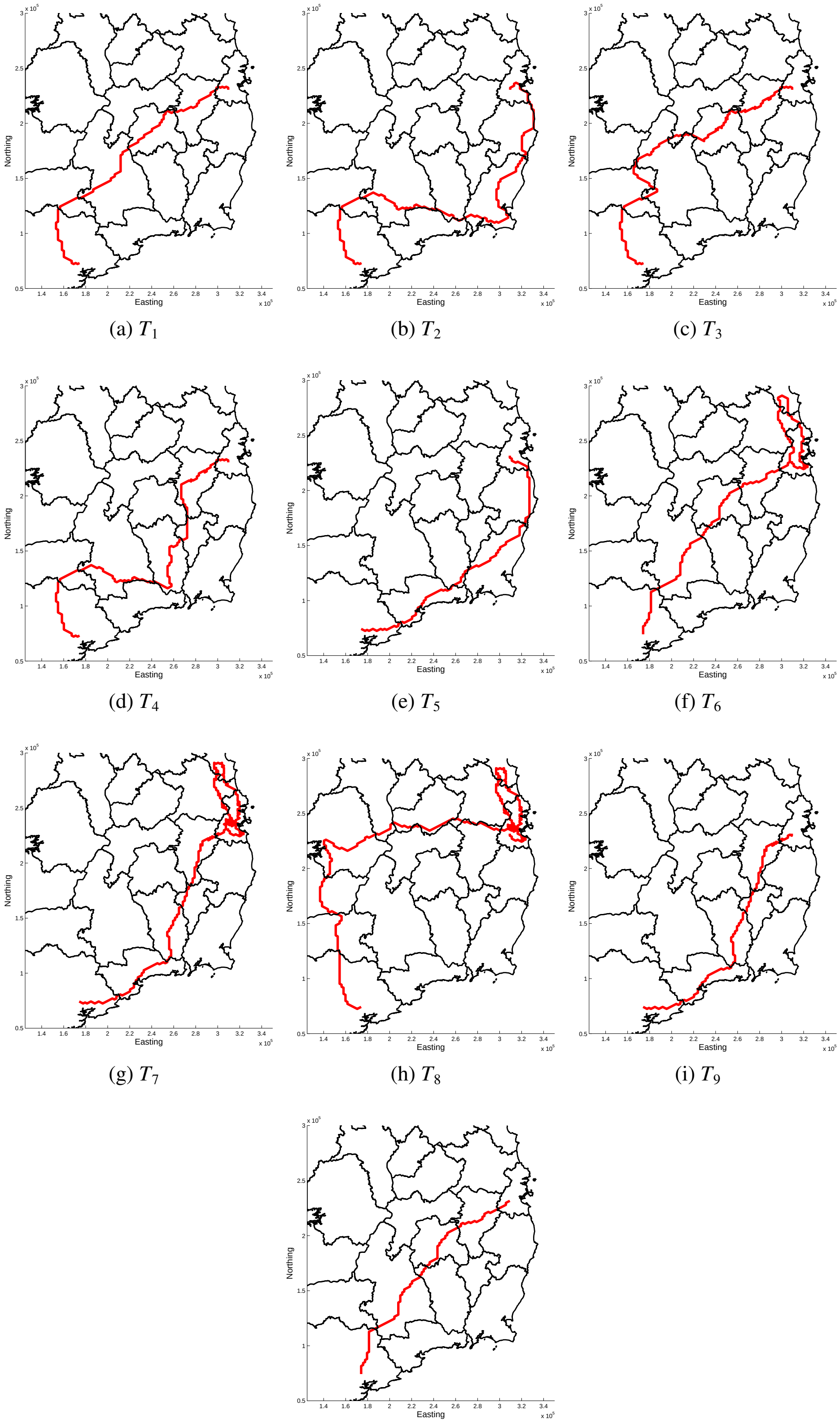
The selection of routes upon which a subscriber may travel while moving between a start and end location may be extracted from Γ by the *k*-shortest path algorithm outlined in Yen [205]. This technique was applied for both rail and road Γ matrices using the MATLAB® function GRAPHKSHORTESTPATHS, where the start and end locations were points within Dublin City and Cork City, respectively. The result was several paths between each point along both rail and road networks, each with varying travel cost. Several of the paths produced were very similar, thus in an effort to reduce the number of paths and to extract the distinct travel routes, clustering was applied based on the similarity of each estimated path. The similarity measurement used was the Hausdorff distance, and clustering was applied using an agglomerative hierarchical based clustering algorithm [179]. The dendrograms illustrated in Figure 4.28, show the clear separation of paths into a few high level clusters. As a representation of each cluster, the path with lowest travel cost within each was selected. The selected paths for rail and road networks are displayed in Figure 4.29. The segregation of journeys between these travel paths are detailed in Section 4.6.1.



(a) Road

(b) Rail

Figure 4.28: Dendrogram illustrating the arrangement of the clusters produced by the agglomerative hierarchical clustering of travel paths from: (a) road travel paths; and (b) rail travel paths.



(a) T_1 (b) T_2 (c) T_3

(d) T_4 (e) T_5 (f) T_6

(g) T_7 (h) T_8 (i) T_9

(j) T_{10}

Figure 4.29: The selected paths from within each cluster with the lowest travel cost for both road and rail networks.

4.6.1 Travel Path Identification

To identify which routes were taken by extracted journey trajectories between Dublin city and Cork city, the similarity between each journey trajectory and estimated travel path was computed using PCC. $D_{PCC}(T, J)$ was then used as a feature to classify the most likely travel path taken as follows:

$$T = \begin{cases} T_x & \text{if } D_{PCC}(T_x, J) \; > \; D_{PCC}(T_y, J) + \epsilon \quad \forall\, y \\ unknown & \text{if } D_{PCC}(T_z, J) \; = \; 0 \qquad\qquad\qquad\quad \forall\, z \\ Indistinguishable & \text{otherwise} \end{cases} \tag{4.28}$$

where $y = 1, \ldots, M_T$, $y \neq \; x$, and $z = 1, \ldots, M_T$, M_T is the number of travel paths being compared and T_x is given as

$$T_x = \underset{i}{\operatorname{argmax}} \, D_{PCC}(T_i, J) \; . \tag{4.29}$$

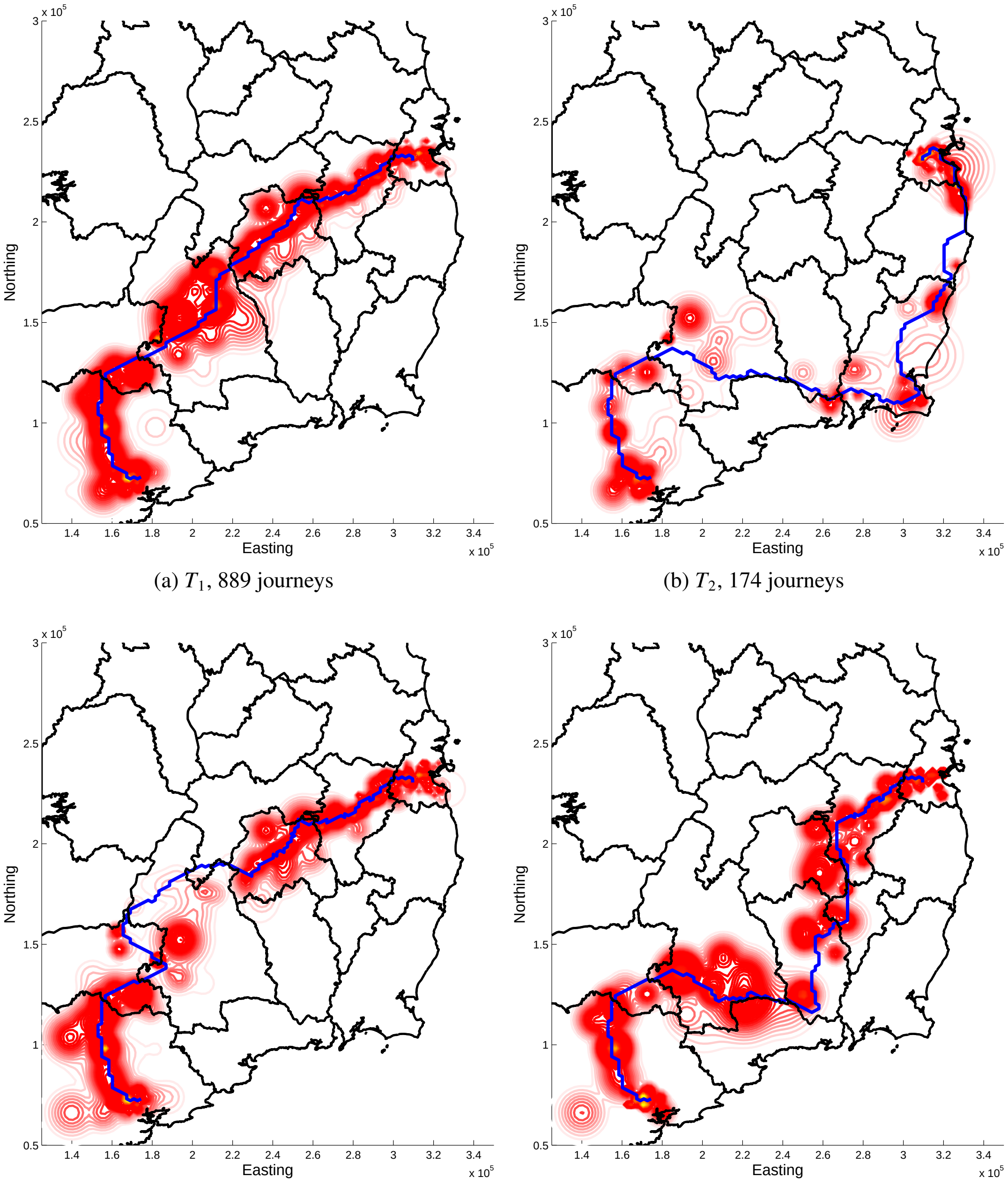
A summary of classification results is given by Table 4.6.

Table 4.6: Classification of travel path taken by 9490 journey trajectories between Dublin City and Cork City.

	Rail				Road							
	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	Ind.	Unknown
No. of J	889	174	749	122	55	1145	258	325	0	815	1087	3871

A kernel density of trajectories assigned to each rail route and road route is given in Figures 4.30 and 4.31, respectively. From visually inspecting the distribution of activity locations related to each classified journey, it is clear that the locations of activities correlates well in most cases with the designated travel routes. Furthermore, as only a limited number of possible travel routes were evaluated, each unique travel path taken by a journey may not have been included in the evaluation process. Thus, there remains 3871 journeys which were not classified as travelling along the tested routes.

The absence of trajectories assigned to T_9 reflects the similarities between T_9 and T_7. This motivates the requirement for a methodology which can discover similar subtrajectory components between J and T similar to that proposed by Buchin *et al* [206]. However, due to time constraints this research is the subject of future work and is not investigated further here.



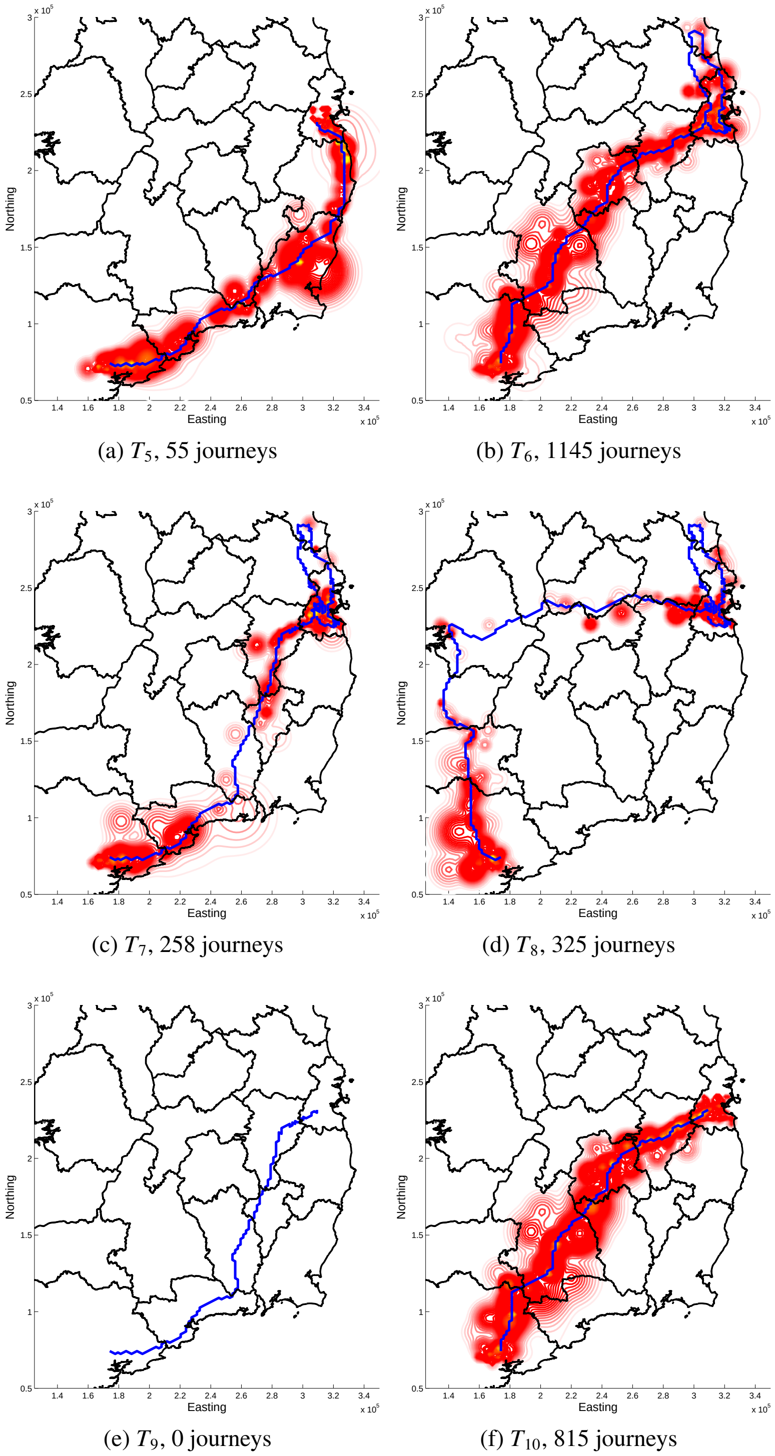
(a) T_1, 889 journeys

(b) T_2, 174 journeys

(c) T_3, 749 journeys

(d) T_4, 122 journeys

Figure 4.30: Kernel density estimate of trajectories assigned to each route on the rail network



(a) T_5, 55 journeys

(b) T_6, 1145 journeys

(c) T_7, 258 journeys

(d) T_8, 325 journeys

(e) T_9, 0 journeys

(f) T_{10}, 815 journeys

Figure 4.31: Kernel density estimate of trajectories assigned to each route on the road network

4.7 Travel Path Generation

Generated travel paths (GTP) represent the likely path taken by a group of similar CDR journey trajectories which move between regions of interest, without prior knowledge of underling transposition infrastructure, or a known travel path. The omission of transportation infrastructural data enables the estimations of travel paths which do not adhere to perceived routes which serve as gateways between each region.

To approximate this path, we find the least cost path between journey end points which is a function of the observed cellular activity of the subscribers who travelled between the regions of interest. The first step in this process is to calculate the total number of activities at each cell tower. Secondly, disperse nodes over the study space and weight each node with interpolated distributed cell tower activity counts. Then, transform node weights into a transition cost matrix Γ (Eq. 4.27), representing the cost of moving from a node to its neighbour. Nodes are selected to form part of the least cost path, if the cost of moving between starting and ending locations while visiting those nodes is minimal in comparison to alternative routes, given only a transition to an adjacent neighbouring node is allowed.

As in Section 4.6, the lattice used comprises of nodes located at the centroids of hexagonal cells. This ensures a constant distance relationship to neighbouring nodes, thus the neighbour transition cost is solely based on Γ. The size of each hexagon (2.5 km in diameter) is chosen to be much smaller than the typical rural area cells (10 km to 20 km). The reason for this choice is to balance the compromise between speed and spatial accuracy. In areas such as the city centre, where cell diameters are sometimes smaller than that of the hexagon grid, the error margin will be the diameter of the hexagon or diameter of the cell, whichever is larger.

Node weights are calculated based on a kernel density smoothing of user activity counts on servicing cell towers. The non-negative spatial kernel density weight W at a node is given by

$$W = \frac{1}{\vartheta} \sum_{i=1}^{N_c} \frac{C_{wi}}{2\pi C_{ri}^2} exp\left(\frac{(x - C_{xi})^2 + (y - C_{yi})^2}{2C_{ri}^2} \right), \tag{4.30}$$

where N_c is the number of cell towers, x and y are the node coordinates and ϑ is a normalisation factor which insures that the largest value of W for all nodes is 1. The Gaussian kernel bandwidth corresponds to cell radius C_r, while C_w represents the number of CDR activities at that cell.

However, kernel density weights cannot be employed directly in a transition cost matrix between neighbouring nodes. The non-negative transition matrix Ω is proposed instead, which relates the observed W at each node to the cost of moving to and from neighbouring nodes.

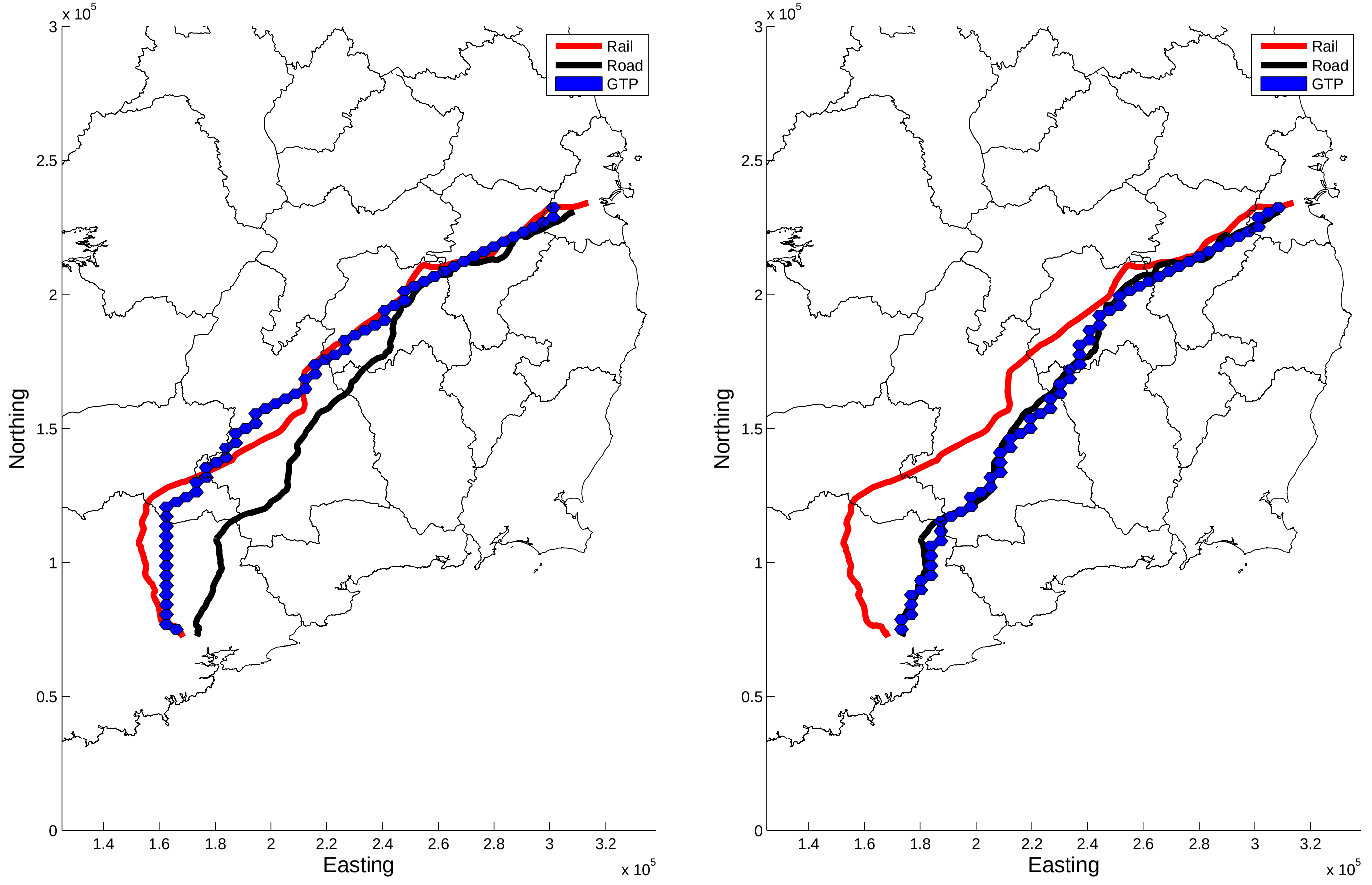
$$\Omega = \begin{pmatrix} \varpi_{1,1} & \varpi_{1,2} & \cdots & \varpi_{1,M_N} \\ \varpi_{2,1} & \varpi_{2,2} & \cdots & \varpi_{2,M_N} \\ \vdots & \vdots & \ddots & \vdots \\ \varpi_{M_N,1} & \varpi_{M_N,2} & \cdots & \varpi_{M_N M_N} \end{pmatrix} \tag{4.31}$$

where M_N is the number of nodes, ϖ_{ij} is the cost of moving between node i and node j and is given by

$$\varpi_{ij} = \frac{W_j}{W_i}\left(2 - W_j\right) \ . \tag{4.32}$$

The ratio W_j/W_i scales the transition cost such that the cost of moving from a high W_i to a low W_j is large, while the cost of moving in the other direction is small. The transition cost of moving from similar weights should also be small, so to penalise the transition among nodes with low W and encourage the shortest path, the term $\left(2 - W_j\right)$ is introduced. Then, to ensure that only a transition to an adjacent neighbouring node is allowed in any one step, each row in Ω is transformed by σ_L, where $\sigma_L(j) = 1$ if node j is a neighbour of node i and $\sigma_L(j) = \infty$ otherwise. We then employ Dijkstra's algorithm [207] to locate the single-source shortest path between starting and ending nodes with non-negative edge path costs corresponding to node transition costs. Dijkstra's algorithm [207] determines the shortest path by selecting the nodes which if visited, return the minimal total transition costs.

An example GTP for the journeys classified as travelling along rail and road transportation link paths is depicted in Figure 4.32. From visual inspection, it can be clearly seen that the GTP corresponds well to the actual rail and road travel paths. It should be noted that each GTP does not exactly match the corresponding transportation link because the CDR locations are locations of cell towers rather than the actual locations of users.



(a) rail journeys

(b) road journeys

Figure 4.32: Generated travel paths of journeys classified as travelling along (a) rail; and (b) road transportation link paths.

4.8 Discussion

This chapter outlined a novel methodology which enables the identification of travel routes taken by subscribers as they move between regions of interest. The observed transitions are summarised using CDR journey trajectories. A journey trajectory is defined as a single recorded path taken between two regions of interest. To preserve privacy, data linking individual subscribers to journeys is removed. The methodology developed uses novel similarity metrics to quantify the similarity between journey trajectories and known travel paths between regions of interest. Using transportation infrastructural point data from OSI [174] a procedure is given which can generate a number of travel paths between start and end locations. These paths are then compared to journey trajectories and likely travel routes assigned. A novel technique to generate a travel path from a group of similar CDR trajectories, without prior knowledge of underling transposition infrastructure or a known travel path, was also developed.

The novel similarity metrics introduced are the virtual cell path distance D_{VCP} and the probabilistic cell connectivity distance D_{PCC}. Temporal extensions were also defined, namely modified virtual cell path distance D_{MVCP} and modified probabilistic cell connectivity distance D_{MPCC}. D_{VCP} is based on the proportion of events which occur at cells that are deemed to represent a route of interest, whereas D_{PCC} is a stochastic distance measurement which calculates the probability of activities within a journey trajectory being along any travel route. Novel enhancements to the longest common subsequence distance, D_{LVCP} and D_{LPCC} were also developed. These enhancements incorporate D_{VCP} and D_{PCC} respectively into the longest common subsequence distance, D_{LCSS}, to enable it to account for CDR trajectory spatial-heteroskedasticity.

The ability of D_{VCP} and D_{PCC} to infer a mobile subscriber's travel path is compared to traditional similarity metrics using a test dataset. The test dataset is comprised of simulated journey trajectories which are generated using a novel agent based model which simulates CDR journey trajectories between regions of interest. Both metrics are shown to outperform traditional techniques when classifying these travel routes.

Using similar simulated journey trajectories, the performance of D_{MVCP}, D_{MPCC}, D_{LCSS}, D_{LVCP} and D_{LPCC} was compared. Results showed that both D_{MVCP} and D_{MPCC} outperformed

D_{LCSS}, D_{LVCP} and D_{LPCC}. Also, incorporating D_{VCP} or D_{PCC} into D_{LCSS} improves the accuracy of D_{LCSS} with respect to CDR trajectory distance calculations. While D_{MVCP} and D_{MPCC} performed best, both D_{LVCP} and D_{LPCC} will naturally account for instances where noise is introduced to a journey trajectory due to outdated cell tower locations. Therefore, when deciding which similarity metric to apply, there is a tradeoff between accuracy and noise tolerance.

The proposed methodology which incorporates these measurements to identify the journey travel path taken by a mobile subscriber between regions of interest is highly accurate when that subscriber travels directly between each region, given required sampling criteria. However, travel path prediction for multipath journeys was less accurate, which is reflected by the number of unknown travel paths recorded in Table 4.6. As a result, the methodology proposed may be better suited to identifying journey travel paths which correspond to direct travel paths. Future work may build upon these results by segmenting each multipath journey into several direct sub-journeys by incorporating a greater number of regions of interest into each study.

Furthermore, due to spatial and temporal sampling issues, the use of CDR billing data for travel path identification is only effective for journeys which cover large distances, and it is not suited to small area studies. Rose [147] reached a similar conclusion, observing that the use of mobile phone sourced data for traffic monitoring may be more suited to an interurban motorway context rather than an urban setting. However, as mobile data usage intensifies due to the introduction of 4G services, such spatial and temporal sampling issues will become less significant enabling small area studies to be carried out.

Also, attention should be given to the customer profile and market penetration of the mobile network operator who supplies the CDR billing information, as it may add bias to results. This bias is introduced because it is common for social demographic groups to be clustered into individual networks. Therefore, mobility patterns observed will reflect the mobility patterns of the operator's subscriber base, and may not be a true reflection of an entire populations mobility.

CHAPTER 5

Population Mobility

A census is the primary tool used by national governments to gather information on population metrics, which includes among others, population count, religious status, marital status and household occupancy. The knowledge obtained dictates future policy on decisions related to the planning of future infrastructure and public services. While the information gathered is extremely important for the delivery of such services, carrying out a census is extremely expensive and time-consuming. As a result, a census may be only carried out every 5-10 years. Consequently, they provide poor temporal resolution and are incapable of providing information on the current status of a population. This motivates the requirement for low cost alternatives.

In this regard, exploiting the ubiquity of mobile devices has become an attractive alternative [67, 68, 69, 70]. While the information sourced from such devices may never replace a census, the population count estimated via techniques such as that discussed by Ahas *et al.* [68] result in much more temporally fine-grained measurements. However, such techniques often require the estimation of mobile subscriber home locations which can be computationally intensive and may have several privacy related issues.

As a result, there is a need for techniques which allow population density estimations which are both computationally efficient and privacy preserving. The research presented within this chapter details novel research into the development and evaluation of such techniques. The first developed technique uses the steady state vector of a modified Markov chain mobility model of individual subscribers to enable their residential locations to be estimated. This is achieved by selecting the region which has the maximum steady state vector weighting. When individual residential location estimates are combined, a measurement of population density can be obtained. The second technique uses the steady state vector of a modified Markov chain mobility model characterising the regional transitions of Meteor's customers, to obtain a direct measure of population density.

Each modified Markov chain mobility model is constructed using Meteor's call detail records (CDR) from 01/12/2010 to 31/01/2011. When compared to data from a recent census held on 10/04/2011, results show that there is a high correlation between estimated population density and census population counts for each of the proposed techniques. An overview of the methodology used in both techniques is depicted in Figure 5.1.

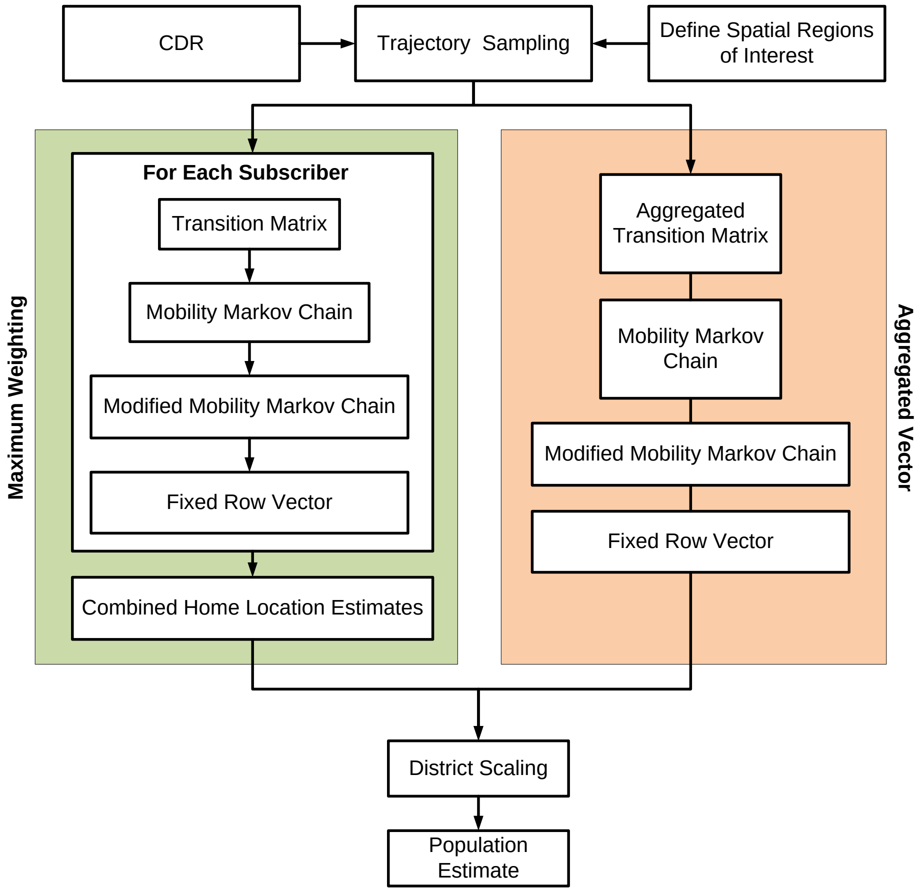


Figure 5.1: Overview of each population estimation technique.

The omission of individual subscriber regions of interest means that the aggregated approach is both privacy preserving and computationally efficient. However, while the use of individual subscriber data reduces the attractiveness of the maximum weighting approach with respect to such criteria, the efficient ranking of each individual's spatial regions of interest in terms of their importance to the individual has both commercial and research applications. These include, among others, geo-marking applications, planning insight and better mobile network optimisations. Some of the methodology required to deliver these applications is further outlined in Chapter 6.

The remainder of this chapter is organised as follows: Section 5.1 gives details on the information which may be extracted from transition matrices that record the flow of subscribers between regions of interest, while Section 5.2 details how such matrices may be transformed into mobility Markov chains. Section 5.3 then discusses how these mobility models may be used to extract measurements of population density comparable to that captured through a census. Section 5.4 then investigates the use of Markov chain mobility model eigenvectors for community identification. Finally, Section 5.5 concludes the chapter with a discussion outlining the benefits and limitations of the proposed techniques.

5.1 Subscriber Transition Intensity

As previously discussed in Section 3.6, the flow of individuals between cells and clustered cell regions may be summarised through an aggregated transition matrix, $\Upsilon_a(k)$,

$$\Upsilon_a(k) = \begin{pmatrix} \upsilon_{1,1}(k) & \upsilon_{1,2}(k) & \cdots & \upsilon_{1,R}(k) \\ \upsilon_{2,1}(k) & \upsilon_{2,2}(k) & \cdots & \upsilon_{2,R}(k) \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{R,1}(k) & \upsilon_{R,2}(k) & \cdots & \upsilon_{RR}(k) \end{pmatrix} \tag{5.1}$$

where R is the number of regions of interest, and $\upsilon_{ij}(k)$ is the transition intensity from region i to j at time k. In this form, the temporal flow of subscribers between cell towers or clustered cell regions is readily available. For example, Figure 5.2 depicts the flow of subscribers into and out of Dublin's city centre during the period 13/12/2010 to 19/12/2010.

As demonstrated by Ahas *et al.* [32], variations in regional occupancy can be quantified using mobile positional data. In an Irish context, variations in regional occupancy can also be observed through $\Upsilon_a(k)$. As a demonstration, Figure 5.3 depicts the influence of University College Dublin (UCD) semesters on the aggregated mobility patterns of the Donnybrook region in Dublin City. UCD is the Republic of Ireland's largest university with approximately 25,000 students. It is located in the Donnybrook region of Dublin city, which has a surrounding resident population of approximately 51,000 [2]. During the observation period, the first academic semester finished on December 17th and the second semester started on January 17th. The increased flow of individuals to and from Donnybrook during each semester is clearly visible in Figure 5.3, highlighting the impact of the university on local area mobility.

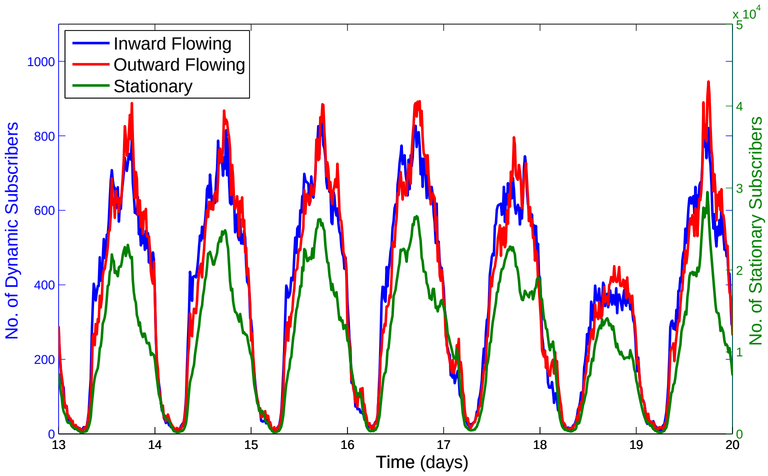


Figure 5.2: The aggregated flow of subscribers to and from a clustered cell region covering Dublin city centre from 13/12/2010 to the 19/12/2010. Inward and outward flow intensity is measured using the left hand axis, while the quantity of observed stationary subscribers is measured using the right hand axis.

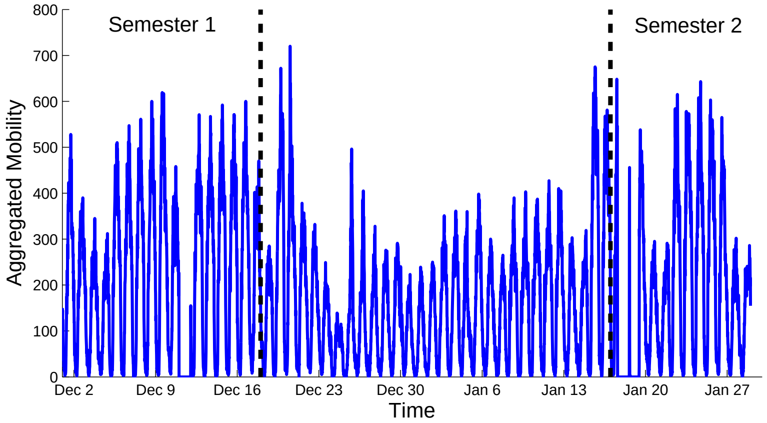
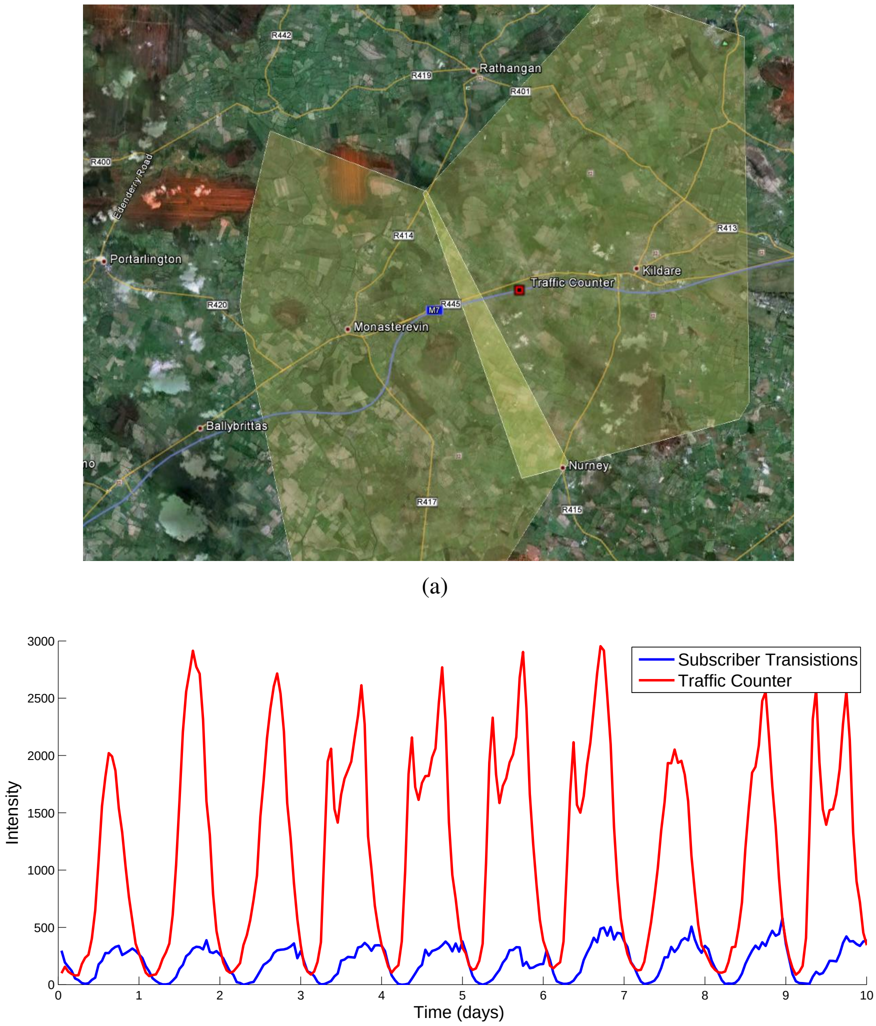


Figure 5.3: The aggregated flow of individuals to and from the Donnybrook region of Dublin City, highlighting the impact of UCD semesters on local mobility patterns.

Such data is useful when trying to observe the temporal flow of people between regions. However, when subscriber transition intensity is compared to traffic counter data, a common measure of flow intensity between regions, there are some notable differences. To demonstrate these differences, the flow of subscribers between clustered cell regions corresponding to the towns of Kildare and Monasterevin were compared to traffic counter data from a counter position on the M7 motorway located between each town. These towns are located in county Kildare and are serviced by both motorway, regional roads and rail connections. The observed transitions between each, as illustrated in Figure 5.4, shows that subscriber transition intensity lags temporally compared to the traffic counter data.



(a)

(b)

Figure 5.4: Comparison of traffic count data and the aggregated bi-directional flow intensity of subscribers moving between Kildare town and Monasterevin: (a) Clustered region coverage areas, town locations and traffic counter location; and (b) Observed transition intensity between each town.

As mentioned to in Section 3.4, this phenomenon occurs because while people are active and travelling early in the morning, for example from 6 AM to 8 AM, they are less likely to be sampled through CDR due to low overall mobile network activity. This motivates the requirement for suitable techniques which can relate the transition intensity observed through CDR to the actual flow of people between regions of interest [150].

5.1.1 Flow Directionality

The temporal directionality of the transition intensity is an important feature when observing the flow of individuals. This is highlighted in Figure 5.5 showing a sample of the average directional movement of subscribers across the Republic of Ireland. This visualisation is constructed using a customised interface between MATLAB® and Google Earth®, where the width of a connecting arrow corresponds to directional flow intensity. Note, very low intensity connection arrows have been removed for visual clarity. The anchor point for each connection arrow is positioned at the centroid of each cluster, which is the centre of mass of cell base station sites within the cluster and is independent of the number of cells in each site. From the figure, inspection the influence of primary roads and motorways on transition patterns between major Irish towns is clearly evident.

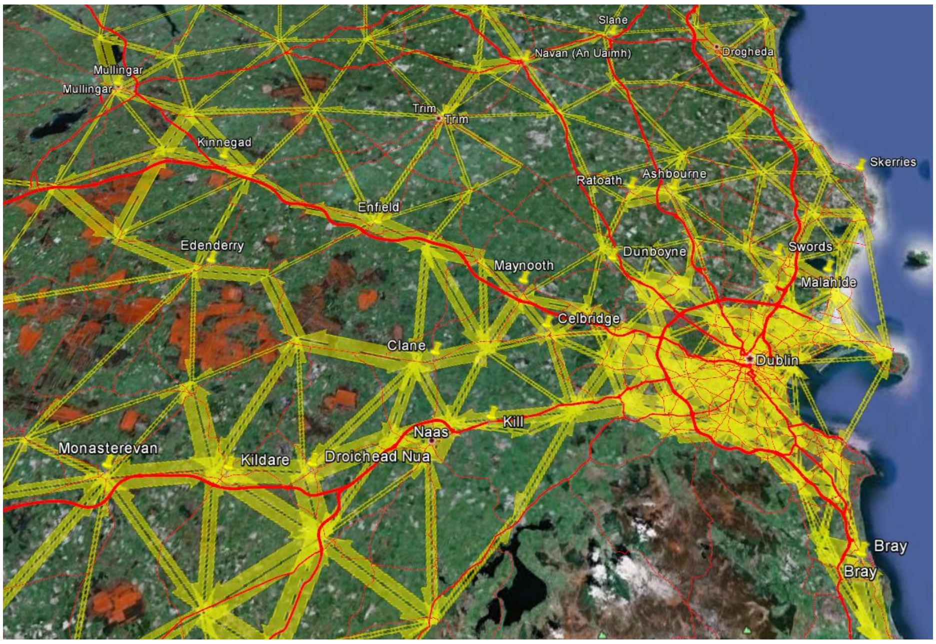
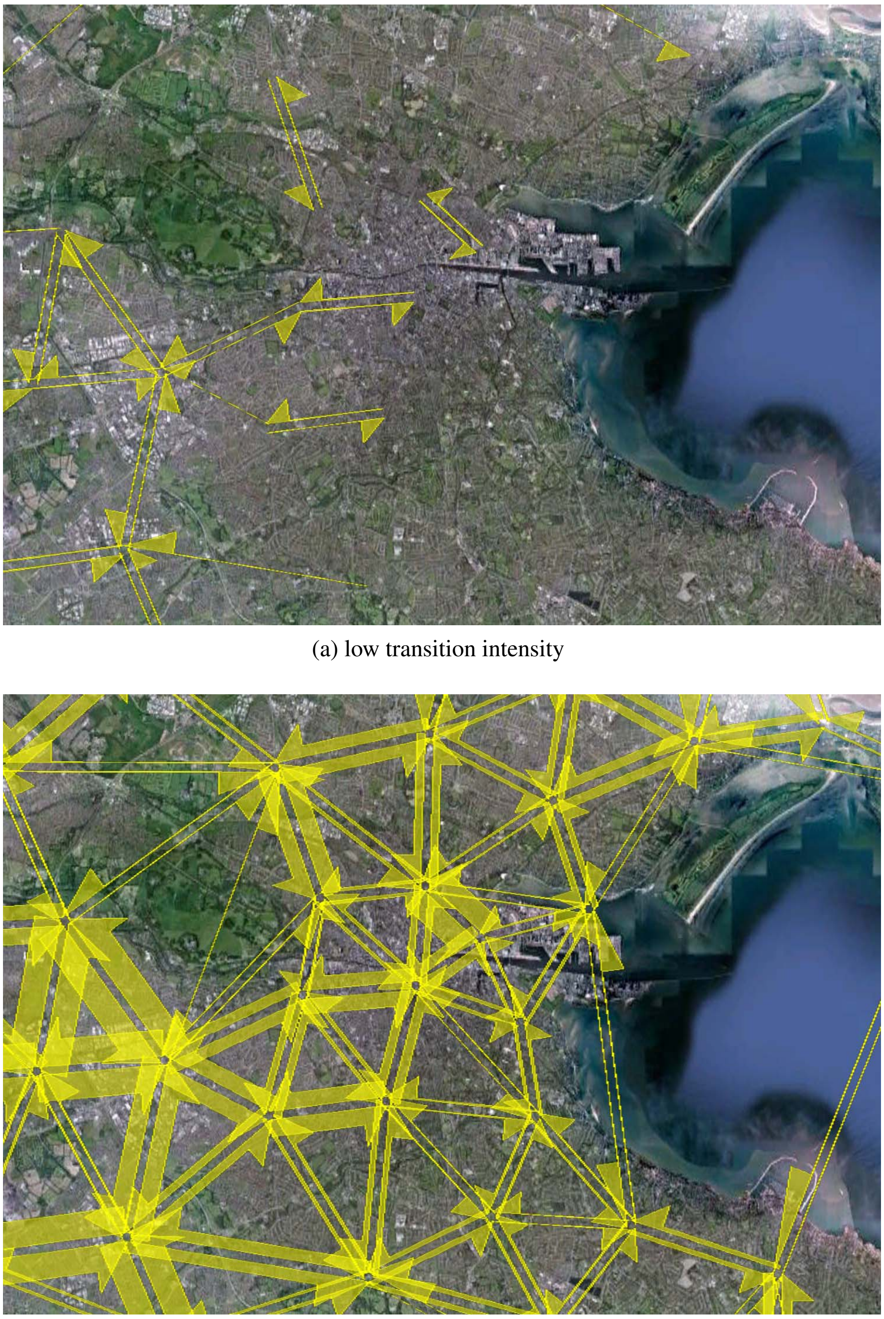


Figure 5.5: Average intensity of subscriber transitions between clustered cell regions.

The temporal component of this transition directionality can also be used to gain further insight into the dynamics governing regional connectivity. To illustrate, Figure 5.6 depicts regional transition flows between cell site clusters across Dublin city for both the time periods of low and high intensity.



(a) low transition intensity

(b) high transition intensity

Figure 5.6: Dublin city regional transition flows in time periods of (a) low; and (b) high intensity, where the width of the connecting arrow corresponds to directional flow intensity. Note connecting arrows with very low intensity have been removed for visual clarity.

5.2 Markov Chain Mobility Model

A Markov chain is a mathematical representation of a stochastic process that undergoes step transitions from one state to another within a finite or countable state space. They have been extensively used in many domain areas including mobility modelling [138, 69], biomedical data analysis [208] and speech recognition [209, 210]. A first-order, discrete-time Markov chain is used to mathematically represent a process, $\{S(k),\ k = 0, 1, 2, \dots\}$, that undergoes random step transitions such that

$$P[S(k) = j \mid S(k-1) = i] = p_{ij}(k) \tag{5.2}$$

for all i,j and k [211]. Here $p_{ij}(k)$ is the conditional probability that the process will transition from state i at time $k-1$ to state j at time k. A Markov chain which does not depend on the time unit, is known as a *homogeneous Markov chain* and implies that

$$P[S(k) = j \mid S(k-1) = i] = p_{ij}. \tag{5.3}$$

From this, it is inferred that the state transition probability p_{ij} only depends on the current state and not on the sequence of previous states. This specific kind of memorylessness is called the Markov property. It is customary to display the state transition probabilities p_{ij} as entries of a $N_s \times N_s$ matrix P,

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N_s} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N_s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N_s,1} & p_{N_s,2} & \cdots & p_{N_sN_s} \end{pmatrix} \tag{5.4}$$

where p_{ij} satisfy the following conditions:

1. $0 \leq p_{ij} \leq 1,\ \forall\, i\ \forall\, j$
2. $\sum_j p_{ij} = 1,\ \forall\, i$, which follows from the fact that the states are mutually exclusive and collectively exhaustive.

A state transition diagram is a convenient way to visually represent the matrix P. States are typically represented by circles and transitions are represented by weighted connections. An example state transition diagram is illustrated in Figure 5.7. It corresponds to a scenario where it was observed via a series of rolls of a particular biased 4-sided die that the outcome of the next roll depends on the outcome of the current roll, as summarised by

$$P = \begin{pmatrix} 0.3333 & 0 & 0.4103 & 0.2564 \\ 0 & 0.1923 & 0 & 0.8077 \\ 0 & 0.3571 & 0.6429 & 0 \\ 0.7059 & 0 & 0.2941 & \end{pmatrix} \tag{5.5}$$

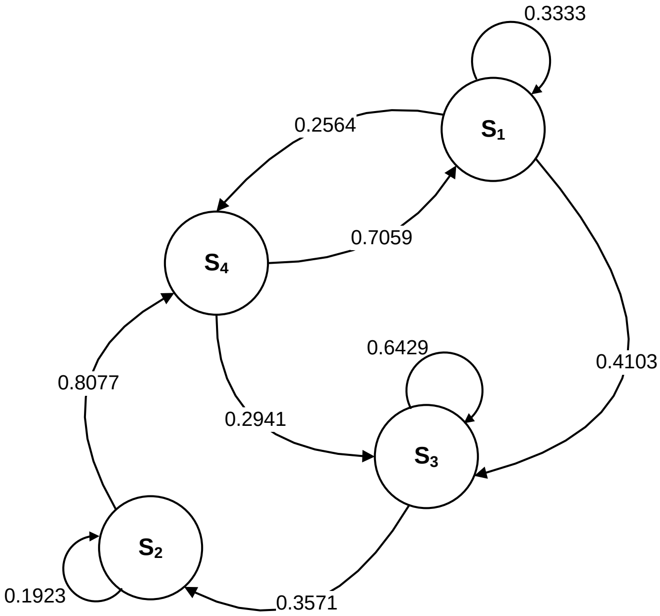


Figure 5.7: A state transition diagram of a particular Markov chain representing the observed outcome from a series of biased 4-sided die rolls.

In general, states of a discrete-time Markov chain may be classified by their transient proprieties. A list of formal state definitions is given as follows [211]:

- A state i is called a transient state if there is a positive probability that the process will never return to i again after it has left i.
- A state i is called a *recurrent* state if, with probability 1, the process will eventually return to i after it has left i.
- A recurrent state j is called a *periodic* state if there exists an integer a (known as the period), $a \geq 1$, such that $p_{ij}(n)$ is zero for all values of n other than a, $2a$, $3a$, $\ldots$; If

$a = 1$, the recurrent state j is said to be aperiodic.

- A recurrent state i is called a *positive recurrent* state if, starting at state i, the expected time until the process returns to state i is finite. Otherwise, the recurrent state is called a *null recurrent* state.

- Positive recurrent, aperiodic states are called ergodic states.

- A chain consisting of ergodic states is called an irreducible, or ergodic, chain.

- A state i is called an *absorbing* state if $p_{ii} = 1$. Thus, when the process enters an absorbing state, it has a 0 probability of leaving, meaning it is absorbed or trapped.

The probability that the process starts in state i and finds itself in state j at the end of the n^{th} transition is given by the product of the probability that the process starts in state i and finds itself in an intermediate state ϕ after r transitions and the probability that it goes from state ϕ to state j after an additional $n - r$ transitions [211]. For all $0 \leq r \leq n$

$$p_{ij}(\phi) = \sum p_{i\phi}(r) p_{\phi j}(n - r) \tag{5.6}$$

From 5.6, it may be shown that $p_{ij}(n)$ is the $(i, j)^{\text{th}}$ entry in the matrix P^n, where P^n is the matrix

$$P^n = \begin{pmatrix} p_{1,1}(n) & p_{1,2}(n) & \cdots & p_{1,N_s}(n) \\ p_{2,1}(n) & p_{2,2}(n) & \cdots & p_{2,N_s}(n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{N_s,1}(n) & p_{N_s,2}(n) & \cdots & p_{N_s N_s}(n) \end{pmatrix} \tag{5.7}$$

If a Markov chain is irreducible, or ergodic, it is possible to go from every state to every other state in one or more steps [212]. If a Markov chain is ergodic, then the following holds true:

$$W = \lim_{n \to \infty} P^n \tag{5.8}$$

where W is a matrix with identical rows w, and all components of w sum to 1. Then $wP = w$, and any row vector v such that $vP = v$ is a constant multiple of w. A row vector w with the property $wP = w$ is called a *fixed row vector* for P and may be calculated by various methods, as outlined in [212]. A fixed row vector characterises the long term probability of a system being in a given state when the state transitions are governed by an underlining Markov chain.

Homogeneous Markov chains are useful when the state sequence, $S(k),\ k = 0, 1, 2, \ldots,$ is directly observable. By extracting a subscriber's CDR trajectory, it is possible to directly observe an individuals cell tower state sequence. As previously discussed, cells may be linked to symbolic locations defined by their coverage regions, thus a Markov chain may be used to model a mobile subscribers transient movements between these symbolic locations, where the number of observable states equals the number of regions of interest N_R. By counting the transitions between clustered regions from concurring activities, an aggregated transition matrix, Υ_u, can be constructed which summarises the movement of the uth subscriber.

To reduce the influence of high frequency transitions and to ensure uniformly sampled trajectories, each subscriber trajectory was sampled at a regular interval every 15 minutes from the start of the observation period. The procedure is illustrated in Figure 5.8. Within each 15-minute temporal window, the estimate of location is based on the last recorded servicing cell tower recorded for that subscriber during that period. When no CDR activity occurs during a temporal window, no sample would be taken.

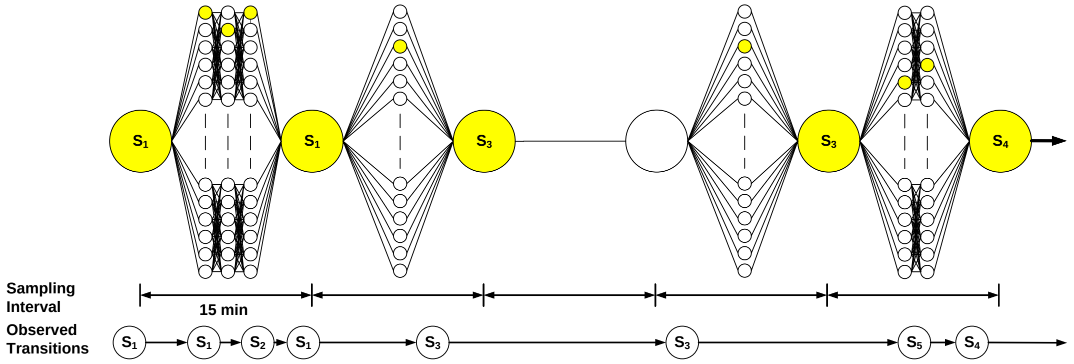


Figure 5.8: CDR trajectory state sequence sampling of the output sequence $S = \{S_1, S_1, S_3, S_3, S_4\}$. Smaller yellow circles represent actual regional transitions within a sample period and larger yellow circles represent the observed output transition sequence before resampling. The larger white circle represents missing information and is discarded.

The transition matrix of each subscriber, Υ_u, can be translated into a transition probability matrix, P_u, by scaling each row such that

$$P_u = [p_{ij}]_{N_R \times N_R} = \sum_{j=1}^{N_R} p_{ij} \;=\; 1\ ,\ \forall\ i \tag{5.9}$$

The resulting transition probability matrix P_u characterises the movements of that individual subscriber between regions of interest. This is illustrated in Figure 5.9 by the state transition diagram for a randomly selected subscriber, where arch height corresponds to transition probability. The visualisation is constructed using customised MATLAB® plotting functions for Google Earth®.

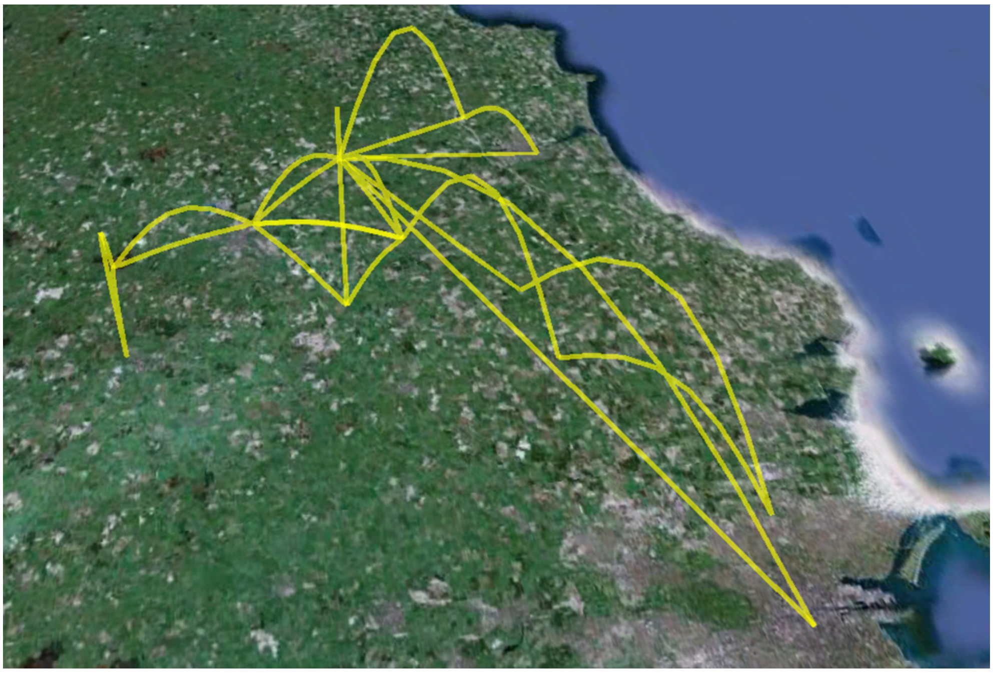


Figure 5.9: Visualisation of a subscriber's transition probability matrix P, where arch heights correspond to transition probability. Note weights which are asymptotically zero are removed for visual clarity.

Similarly, national mobility can be modelled when subscriber movements are combined into a single mobility model characterising flow throughout the country.

$$\Upsilon = \sum_{u=1}^{N_u} \Upsilon_u \tag{5.10}$$

where N_u is the number of subscribers. As before, the aggregated transition matrix, Υ, can be translated into an aggregated transition probability matrix by scaling each row such that

$$P = [p_{ij}]_{N_R \times N_R} = \sum_{j=1}^{N_R} p_{ij} \ = \ 1 \ , \ \forall \, i \ . \tag{5.11}$$

An example aggregated Markov chain mobility model which characterises the flow of individuals between clustered cell regions across the Republic of Ireland is given by the state transition diagram depicted in Figure 5.10. Here, cells have been grouped into distinct clusters as detained in Section 3.5 where $N_R = 500$. For visual clarity, p_{ij} in each instance has been modified such that $p_{ii} = 0 \; \forall \; i$. p_{ij} is then re-normalised as in Equation (5.11). The visualisation is constructed using MATLAB® plotting functions. Note that the opacity of each observed connection edge is dictated by transition probability, p_{ij}.

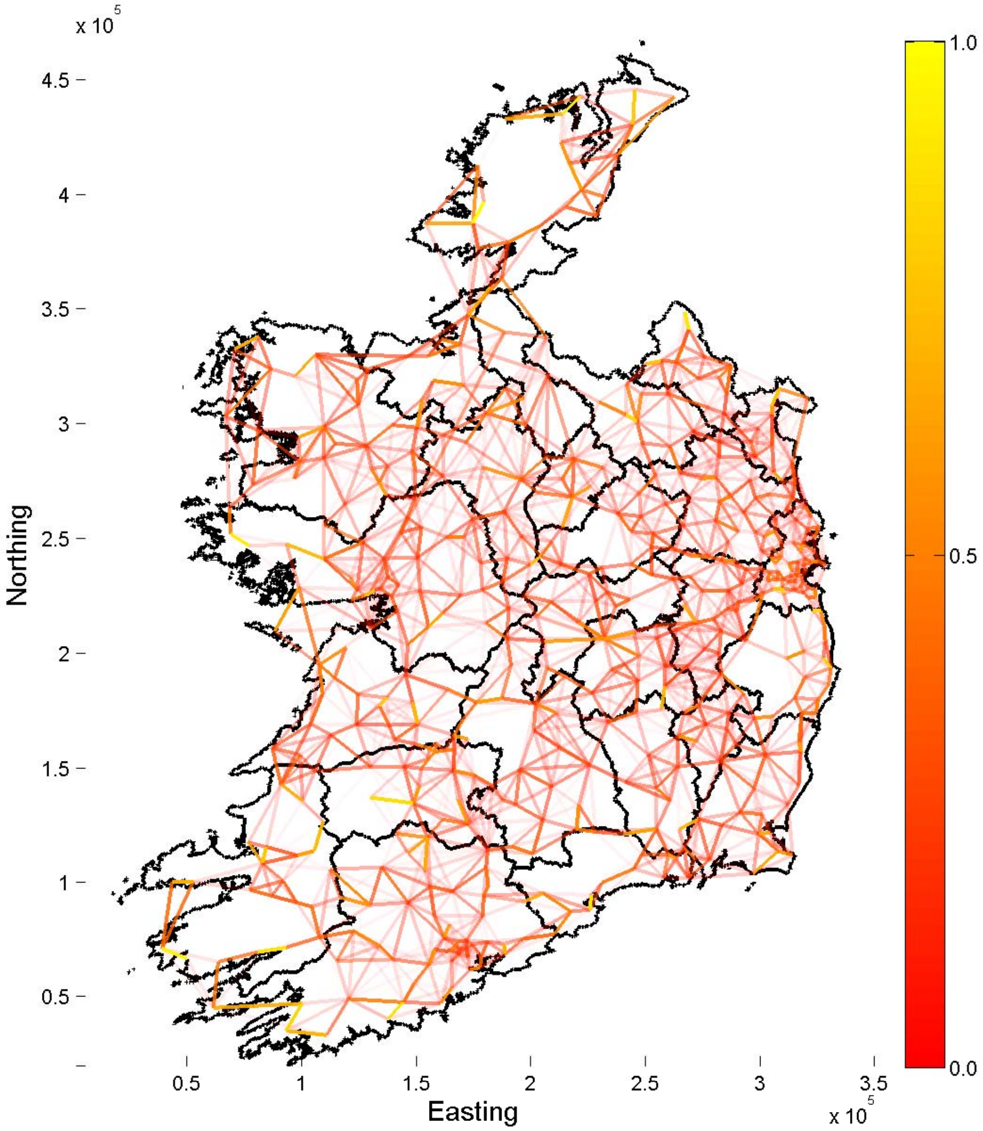


Figure 5.10: Visualisation of aggregated probability matrix, P, characterising the flow of individuals across the Republic of Ireland, where line colour corresponds to transition probability as shown on colour bar.

5.3 Population Density Estimation

As previously discussed, if a Markov chain is ergodic, then it is possible to find a fixed row vector w with the property $wP = w$ [212]. This vector characterises the long term probability of a system being in a given state when the state transitions are governed by an underlining Markov chain. The fixed row vector of a mobile subscriber's mobility Markov chain, w_u, conveys the probability of observing that subscriber at a region in space over a long period of time, if their Markov chain is stationary. In the context of this research, each subscriber's mobility Markov chain is assumed to be stationary over the study period. In order to extract national population counts using fixed row vectors, the home location of each subscriber needs to be segregated from these regions of interest. Here, the maximum weighting approach involves assigning a subscriber's residential location to the region that has the maximum fixed row vector weight. The population count of any region may then be calculated by counting the number of subscribers who are estimated to live in that region.

Alternatively, it is also hypothesised that the fixed row vector for the aggregated Markov chain mobility model, w_a, will convey the likelihood of observing the mobile operators active subscriber base at a particular region in space over a long period of time, which in turn provides an estimate for national population density. Similarly, it is assumed that the aggregated Markov chain mobility model is stationary. The model has the advantages that the calculation is based on the overall subscriber data rather than individual subscriber regions of interest. Hence, it is totally privacy preserving in the sense that none of the subscribers are individually tracked. Also, only a single calculation is required to form the aggregated model fixed row vector, which is less computationally intensive than the maximum weighting approach, where the number of calculations is proportional to the number of subscribers.

However, mobility Markov chains are not necessarily ergodic. Instead, they are typically sparse and may contain absorbing states (i.e. $p_{ii} = 1$). These may occur if a subscriber is only ever serviced by a single cell tower or if its last trajectory sample was to a previously unvisited tower. For an aggregated mobility Markov chain, an absorbing state may occur if subscribers from a particular region of interest never left that area during the time period concerned. In other words, a non-ergodic chain may form if every region of interest was not visited during the observation period.

To ensure each mobility Markov chain is ergodic and thereby non-absorbing, a regularisation process similar to that used by the Google® PageRank algorithm [213] is introduced. It consists of applying a small transition weight to all state transitions before the fixed row vector is calculated and is given by

$$Q = \alpha P + (1 - \alpha)\frac{Z}{N_R} \tag{5.12}$$

where Q is a modified Markov chain, Z is an $N_R \times N_R$ matrix of ones and α balances the learnt mobility patterns summarised by P with the influence of random transition probabilities introduced by the term Z/N_R. To this end α is estimated as $(1 - 1/N_R)$. Note, Q should comply to the following conditions

1. $p_{ii} < 1, \ \forall \ i$
2. $0 \leq p_{ij} \leq 1, \forall \ i, \forall \ j$
3. $Q = [q_{ij}]_{N_R \times N_R} \rightarrow \sum_{j=1}^{N_R} q_{ij} = 1, \ \forall \ i$

The incorporation of uniformly regularised weighting has the added benefit of accounting for the likelihood of observing transitions which relate to all plausible but unobserved journeys. Using previously mentioned clustered cell regions ($N_R = 500$) as a proxy for spatial regions of interest, a visualisation of Q for the selected subscriber whose Markov chain mobility model is depicted in Figure 5.9 is illustrated in Figure 5.11. The observed regional weighting suggests that the subscriber tends to travel in County Meath, with occasional trips into Dublin City.

Using the same clustered regions, an estimate of population density was calculated using both proposed techniques. The results are visualised by Figure 5.12, with density weights normalised between 0 and 1 for visual clarity. While maximum weighting relies only on information collected from individual subscribers, it is prone to noise in CDR data as it relies on the assumption that both a significant amount of time is spent and a significant amount of CDR activities are carried out at residential locations by each subscriber, which may not be true. The mobility Markov chain is constructed such that it takes account of both aspects of user behaviour and reflects that in the form of individual fixed row vectors. Comparing Figure 5.12 with the locations of towns and urban districts in the Republic of Ireland as presented in Figure 5.13, it can be seen that each area of high proportional population density generally corresponds well to urban centres and large towns.

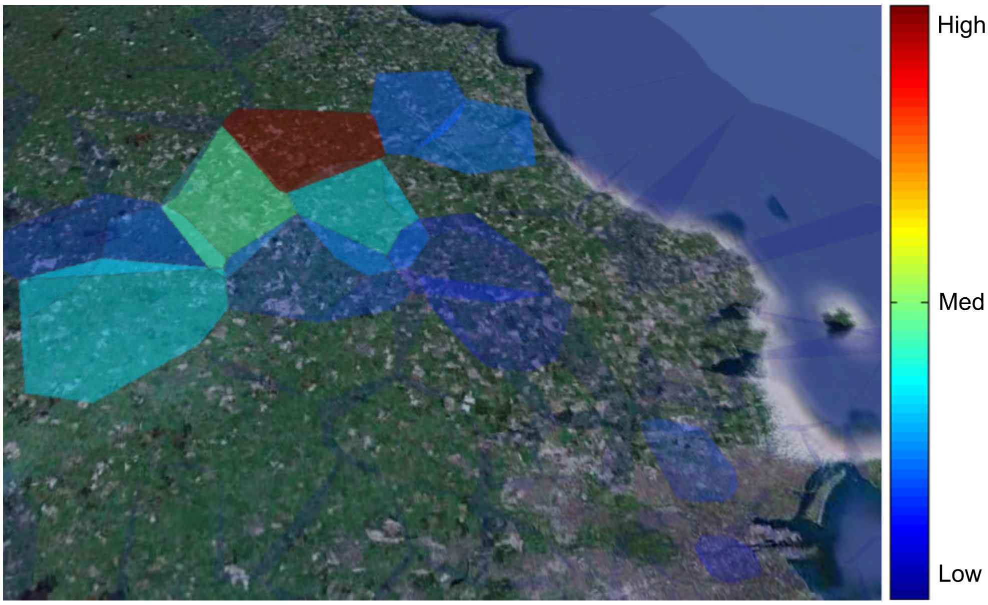


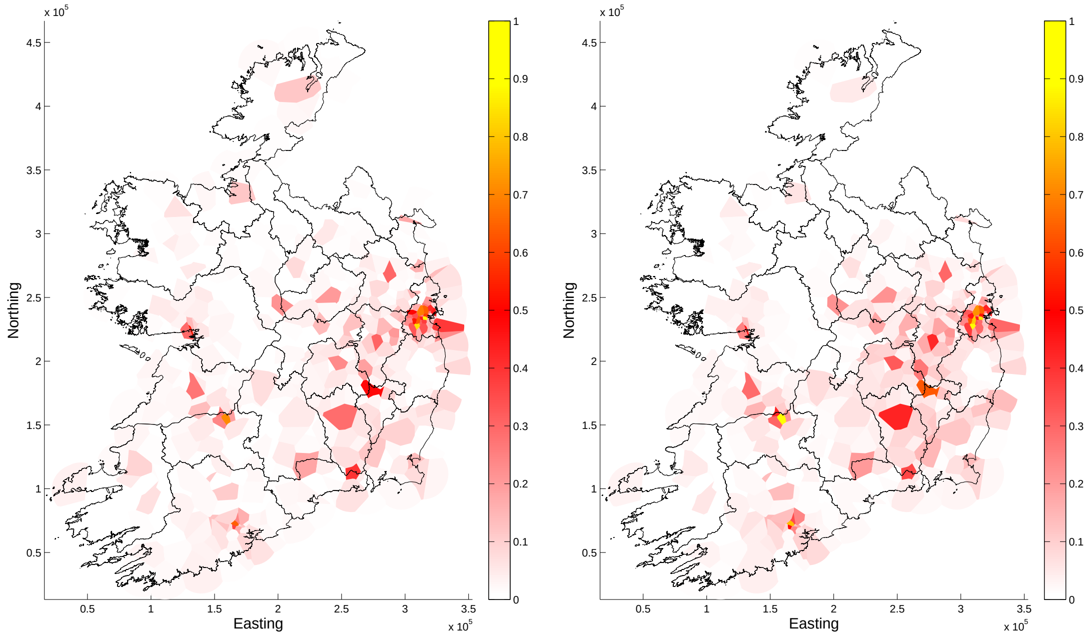
Figure 5.11: Visualisation of the fixed row vector, w_u, for the modified mobility Markov chain, Q_u, of the subscriber whose P is depicted in Figure 5.9.

5.3.1 District Scaling

Problems arise with the estimation of population density through CDR as both clustered cell regions and cell coverage areas do not naturally correspond to the boundaries of districts or municipalities used by governments in the calculation of regional or local population. To allow direct comparisons between estimated population density and census ground truth, where census data is supplied from the Central Statistics Office Ireland (CSO) [2], measurements of population observed at each region need to be redistributed to regions from officially defined district boundaries. In Ireland, a common local area district used in the calculation of population is known as an Electoral Division (ED). There are approximately 3400 ED in Ireland ranging in size from several hundred meters squared in urban areas to several squared kilometres in rural regions.

A sample of the spatial distribution of buildings is displayed in Figure 5.14. The property usage and location of each building is sourced from Geodirectory [214]. Established and maintained by An Post and Ordnance Survey Ireland (OSI), it is one of the most comprehensive building address databases available in the Republic of Ireland.

The procedure used to distribute estimated populating density through fixed row vector analysis onto EDs consists of several steps. First, assign each identified occupied home from



(a) Maximum weighting

(b) Aggregated vector

Figure 5.12: Population density estimates based on (a) individual home locations as sourced from subscribers; and (b) aggregated mobility model.

Geodirectory to an ED. Next, allocate each home in an individual ED to the cell region of interest whose spatial coverage polygon covers that dwelling. If multiple regions of interest cover a particular building, due to instances of overlapping cell tower coverage, randomly assign a covering region from that list. Once all dwellings have been assigned, group them into a matrix H,

$$H = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N_R} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N_R} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M,1} & h_{M,2} & \cdots & h_{MN_R} \end{pmatrix} \tag{5.13}$$

where h_{ij} is the number of homes from ED i assigned to region of interest j, and M and N_R are the number of EDs and clustered regions, respectively. H is then ED normalised such that

$$\bar{H} = [\bar{h}_{ij}]_{M\times N_R} \rightarrow \sum_{i=1}^{M} \bar{h}_{ij} = 1 \;\forall j. \tag{5.14}$$

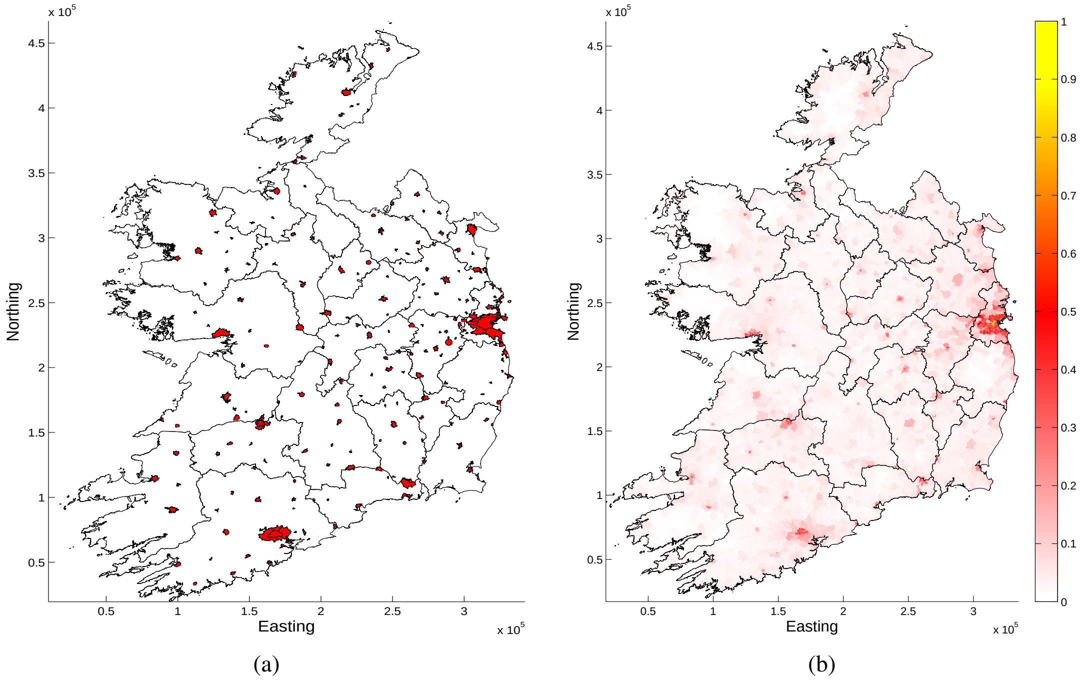
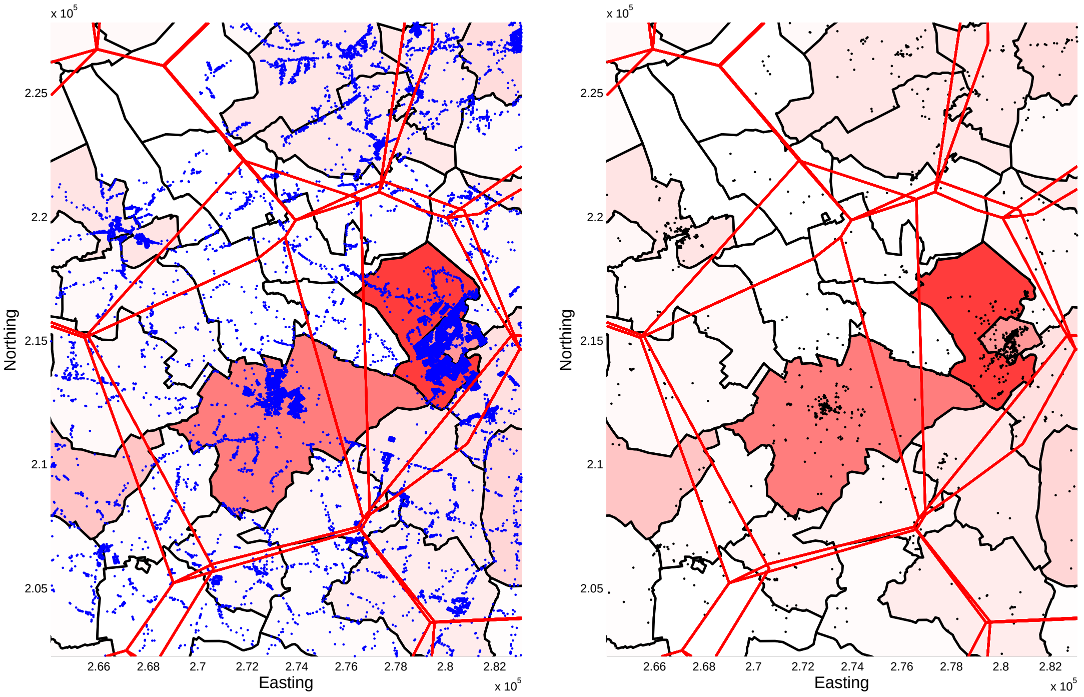


Figure 5.13: Sourced from the Central Statistic Office (CSO) Ireland, (a) town locations across the Republic of Ireland, with (b) corresponding normalized population density.

The number of subscribers living within an individual ED i, $\bar{N}_i$, is then estimated as,

$$\bar{N}_i = \sum_{j=1}^{N_R} N_j \bar{h}_{i,j} \ . \tag{5.15}$$

where N_j denotes the number of estimated subscribers living in a region of interest. Using this method of distribution, Figure 5.15 depicts the population for each ED as estimated using the aforementioned fixed row vectors techniques for $N_R = 500$. In particular, proportional population count estimated for the Dublin region is displayed in Figure 5.16. It can be observed that the ED segregated spatial distribution of subscribers between census data and both estimation techniques are strongly correlated. The discrepancies, such as in the city centre region, could be attributed to the differences in the nature of census, where only residential addresses are recorded, and the human activity observed via mobile networks.



(a) Residential

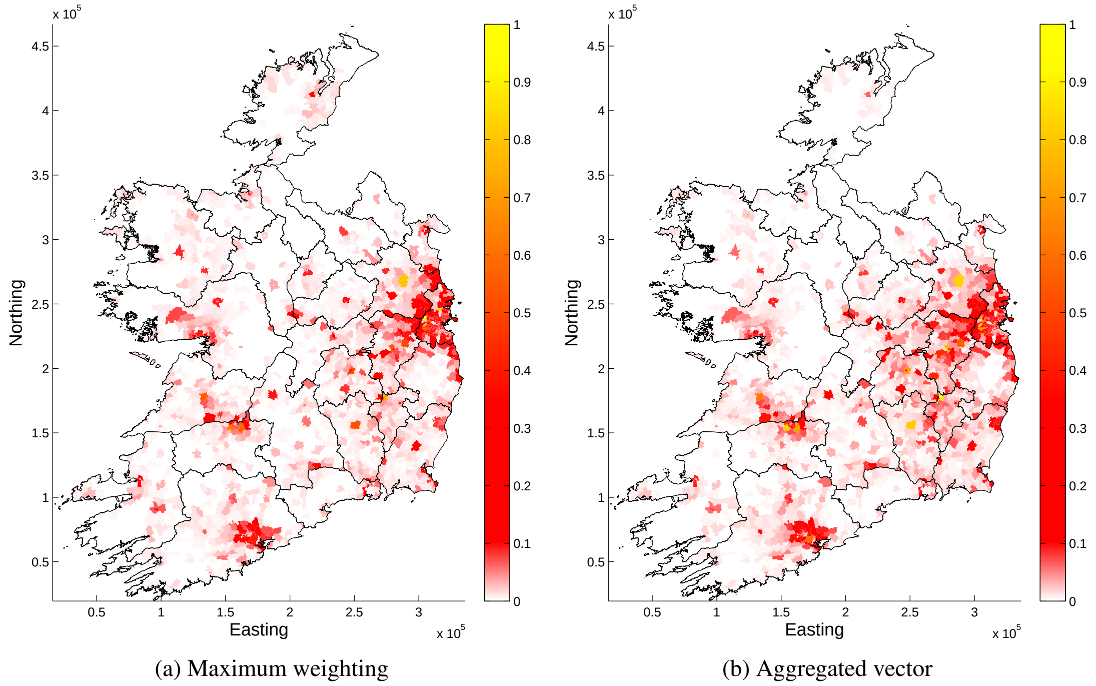
(b) Commercial

Figure 5.14: A sample of the spatial distribution of buildings across the Republic of Ireland. Also included is the cell coverage regions in the area (indicated by red lines), and ED boundaries (black lines); (a) Residential locations (blue dots); and (b) Commercial buildings (black dots).

5.3.2 Census Validation

To validate the population estimate given by each fixed row vector from modified Markov chain mobility models, a direct comparison is drawn between the estimated populations and the Irish 2011 census (CSO, 2012). The correlation of census population counts with the population estimates based on maximum weighting was found to be 0.8645 while that with aggregated vector was 0.8088. The results indicate that both approaches have a strong spatial relationship to census count measurements.

On a national level, the spatial variance of percentage error between census data and estimated population is shown in Figure 5.17. In general, the mean of the percentage error between census data and estimates from aggregated vector is 0.64037% with a standard deviation of 0.51335% while the corresponding values for the population estimates based on maximum weighting are 0.54007% and 0.42464%, respectively. As a result, the maximum weighting approach appears to provide population estimates which match more closely with

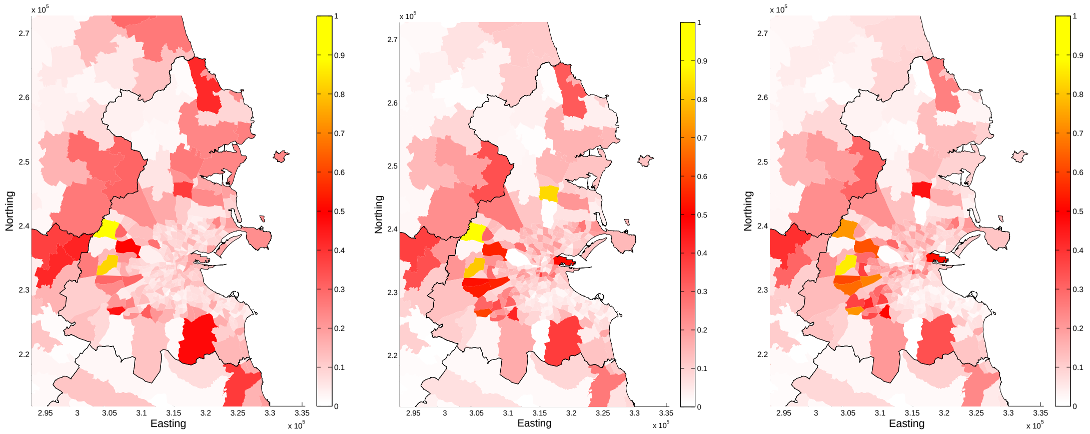


(a) Maximum weighting

(b) Aggregated vector

Figure 5.15: Electoral division population estimates across the Republic of Ireland from (a) maximum weighting; and (b) aggregated vector.

the census data. Note the percentage error is calculated based on normalised population count. From Figure 5.17, there is no clear pattern associated with the spatial distribution of error. In the absence of accurate Meteor subscriber demographics, it is hypothesised that estimation error fluctuates with the spatial density of Meteor's subscriber population. If the age of each subscriber was known, this hypothesis could be tested by proportionally scaling each population estimate by its corresponding ED age profile.



(a) Census (b) Maximum weighting (c) Aggregated vector

Figure 5.16: Proportional population estimated for the Dublin region from (a) census counts; (b) maximum weighting; and (c) aggregated vector.

Finally, to obtain the population density, perhaps a more important measure, H is transformed to D where

$$D = H\mathring{A}. \tag{5.16}$$

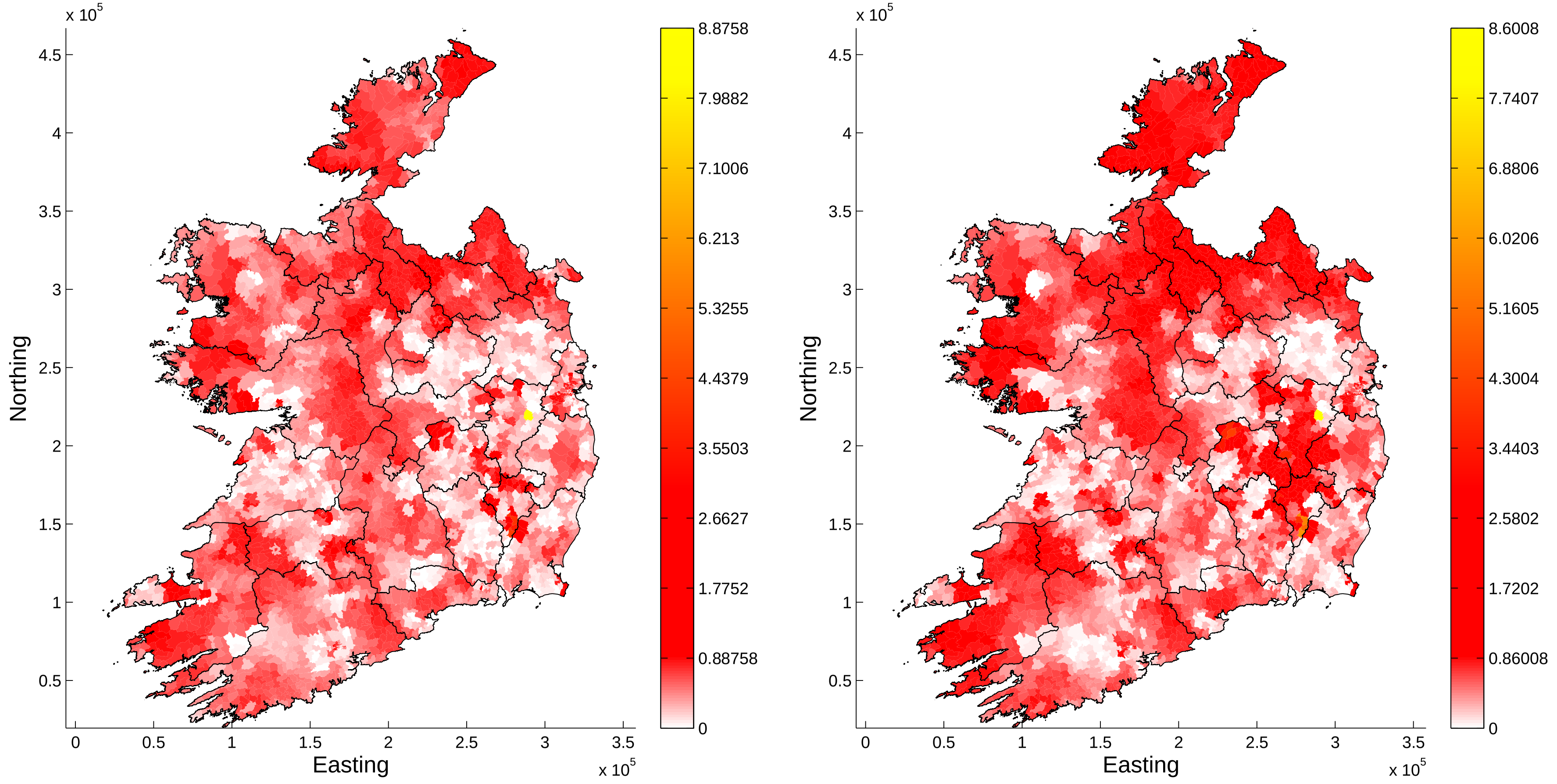
and $\mathring{A}$ is a diagonal matrix

$$\mathring{A} = \begin{pmatrix} \frac{1}{\mathring{a}_1} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{\mathring{a}_2} & 0 & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{1}{\mathring{a}_M} \end{pmatrix} \tag{5.17}$$

where $\mathring{a}_i$ is the spatial area of ED_i. Then, D is normalised to $\hat{D}$ using ED columns similar to $\bar{H}$.

Comparing the population density between census data and both techniques, correlations of 0.8661 and 0.8438 are obtained from maximum weighting and aggregated vector approaches, respectively. While the census correlation from $\hat{H}$ are similar to those from $\hat{D}$ using both techniques, it appears that the maximum weighting approach provides better estimates compared to aggregated vector. However, it is noted that maximum weighting is much more computationally intensive and privacy non-preserving as the assumed home location of each individual subscriber is tracked anonymously during the process.

On the regional scope, a comparison of census data and estimated proportional population



(a) Maximum weighting

(b) Aggregated vector

Figure 5.17: The spatial variance of percentage error between census data and estimated population from (a) maximum weighting; and (b) aggregated vector. note the percentage error is calculated based on normalised population count.

count for each county in the Republic of Ireland is summarised in Table 5.1. In each case, the total proportional population count for each county is the sum of all ED counts which are located in a particular county. From this table, the maximum weighting approach and aggregated vector approach had a percentage mean squared error (MSE) with CSO census of 6.2288 and 4.0415, respectively. When the measurement for Dublin county is omitted the percentage MSE was 1.0491 and 2.0832, respectively.

In the study of mobility, correlation is another important measure, as it captures the relationship between measured population count and population estimates. By observing changes in correlation, we can capture relative displacements of population over time. Here, the maximum weighting approach and aggregated vector approach had correlations with CSO census of 0.98408 and 0.97731, respectively. When the measurement for Dublin county is omitted, the correlations were 0.9124 and 0.8515, respectively.

The strong correlations (> 0.95) of estimated counts demonstrates the effectiveness of Markov chain fixed row vector analysis for approximating proportional population density based on CDR at county level. However, while similar correlations are also observed at the Electoral Division level, there is an overall reduction in the measured correlation at ED relative to county level. This may be a result of greater fluctuations in proportional population representation over smaller geographical regions, caused by the spatial variations of mobile operator penetration.

Moreover, the consistently high levels of errors found in Dublin County might illustrate one possible limitation of a census, that it only contains records of residential addresses. Indeed, subscribers may tend to carry out a substantial amount of CDR activities in densely populated areas. Alternatively, a substantial number of subscribers might be staying in the city centre temporarily for an extended period of time (shopping, night-out, etc), and may not be recorded as residing there through a census.

Evaluating the impact that spatial resolution has on accuracy and correlation with respect to census measurements of ED population density and population count, it was observed that the aggregated approach had an optimal region of spatial resolution. To illustrate Figures 5.18 and 5.19 depict the observed measurements of correlation and MSE over a range of R, a parameter that generally corresponds to spatial resolution.

Table 5.1: A comparison of census data and estimated population density for each county in the Republic of Ireland, where measurements are the percentage of total population.

County	**Central Statistics Office Ireland %**	**Maximum Weighting %**	**Aggregated Vector%**
Carlow	1.19	2.22	2.93
Cavan	1.60	0.53	0.39
Clare	2.55	2.81	3.16
Cork	11.31	10.24	11.98
Donegal	3.51	1.14	0.48
Dublin	27.75	39.40	35.03
Galway	5.46	5.27	4.17
Kerry	3.17	1.36	0.92
Kildare	4.58	6.09	7.24
Kilkenny	2.08	2.24	2.98
Laois	1.76	2.35	3.33
Leitrim	0.69	0.14	0.07
Limerick	4.18	5.36	6.82
Longford	0.85	0.50	0.44
Louth	2.68	1.34	0.70
Mayo	2.85	1.35	0.99
Meath	4.01	3.74	3.55
Monaghan	1.32	0.27	0.15
Offaly	1.67	1.43	1.83
Roscommon	1.40	0.66	0.52
Sligo	1.43	0.65	0.34
Tipperary	3.46	2.02	2.37
Waterford	2.48	1.97	1.69
Westmeath	1.88	1.89	2.19
Wexford	3.17	2.63	3.14
Wicklow	2.98	2.39	2.62
MSE	**0**	**6.2288**	**4.0415**
MSE Excluding Dublin	**0**	**1.0491**	**2.0832**

Results indicate that if spatial regions of interest are too small, the approximation error due to noise, such as those introduced by localisation uncertainty and network penetration, will increase. This suggests that spatial regions of interest should mirror national population centres, where towns and other urban district boundaries are maintained. Note, the effects of spatial resolution on subscriber based maximum weighting could not be evaluated in this extent due to computational limitations. Nonetheless, results presented by Eagle *et al.* [69, 70] show that when locating the home of a subscriber, a set of home cell towers is typically found. Thus, similar to the results obtained for the aggregated vector approach, it is reasonable to assume that the subscriber maximum weighting approach will also have a threshold of R above which transition noise will affect home location estimation accuracy.

Another important factor when evaluating performance is temporal homogeneity. If each mobility Markov chain used in the calculation of population is temporally heterogeneous, it is statistically different when computed for different observation periods, meaning each estimate of population will be different. Thus, the accuracy of the proposed technique will be a function of the observation period. However, due to the limited amount of available data and time constraints, the study of temporal homogeneity is the subject of future work.

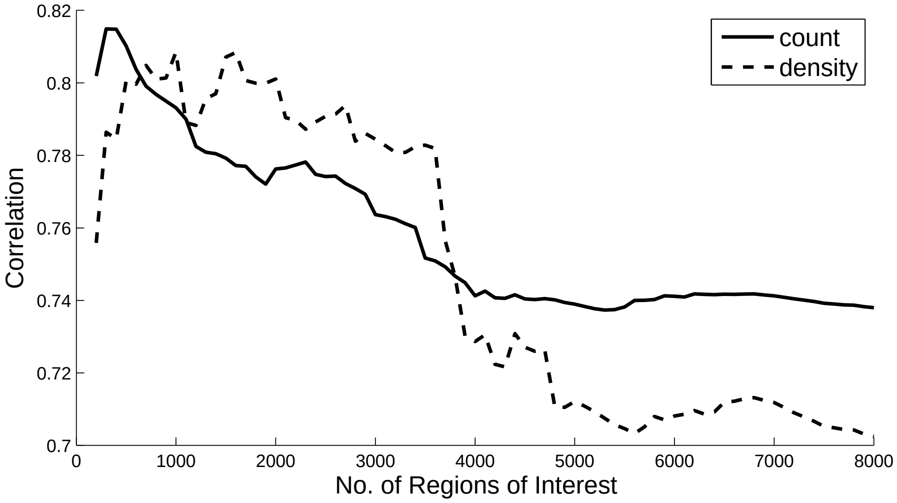


Figure 5.18: Impact of spatial resolution on the correlation between census data and population density estimates form the aggregated vector approach.

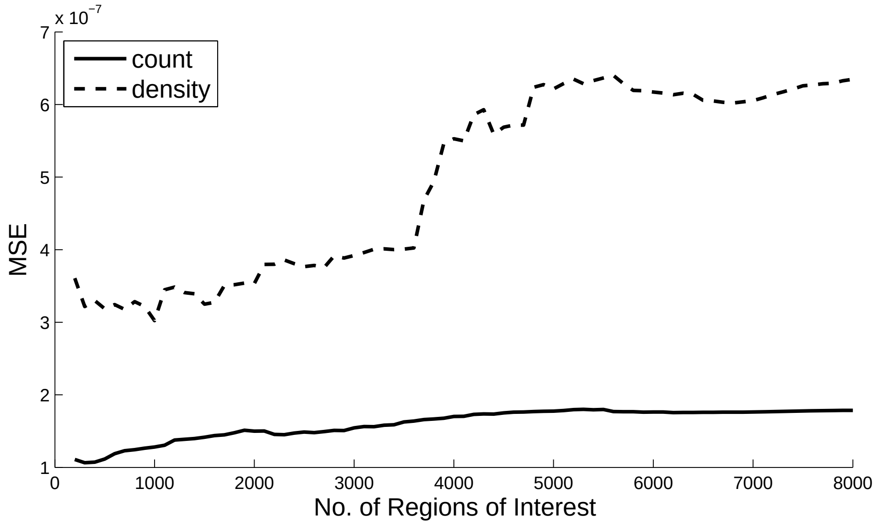


Figure 5.19: Impact of spatial resolution on the MSE between census data and population density estimates form the aggregated vector approach.

5.4 Community Structures

Recent studies by Schlote *et al.* [215] have demonstrated that if the eigenvector of the second largest eigenvalue of a Markov chain mobility model is real, it can be used to characterise hidden communities within an urban network. Applying a similar methodology as Schlote *et al.* [215], the eigenvector relating to the second largest eigenvalue of the aggregated mobility Markov chain model (w_2) was used to evaluate the community structures that exist in the Republic of Ireland. The spatial structures of identified communities across the Republic of Ireland are depicted in Figure 5.20. As expected, identified communities are concentrated around dense urban areas with close links to those regions that are typically associated with commuting.

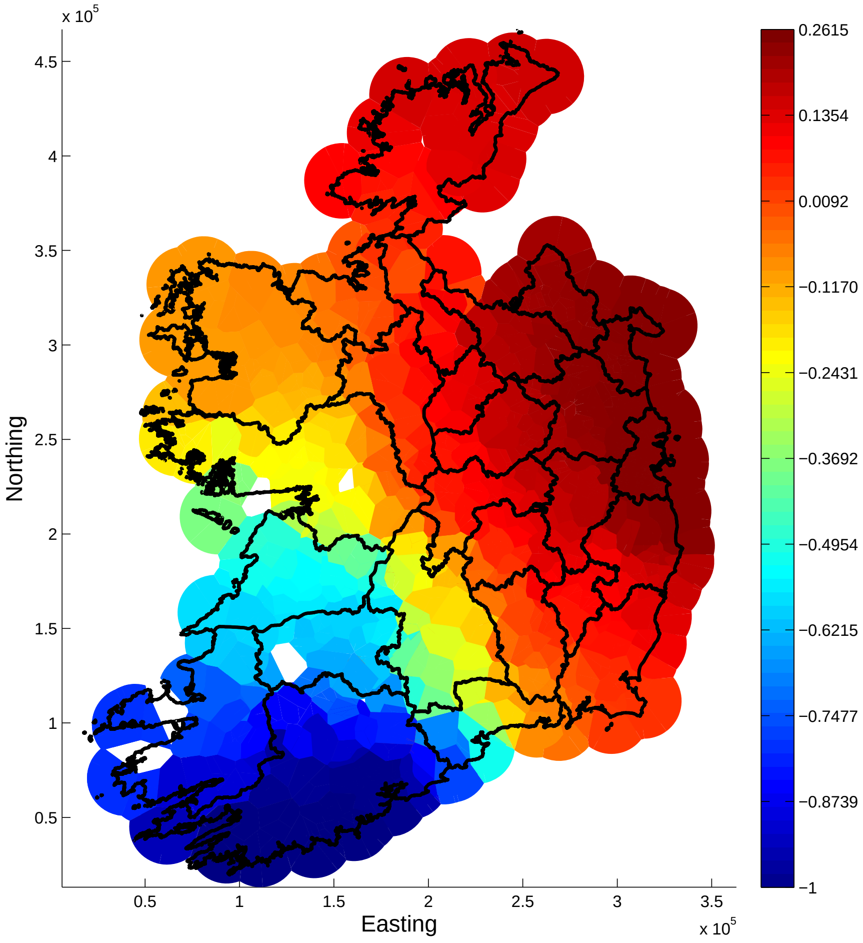
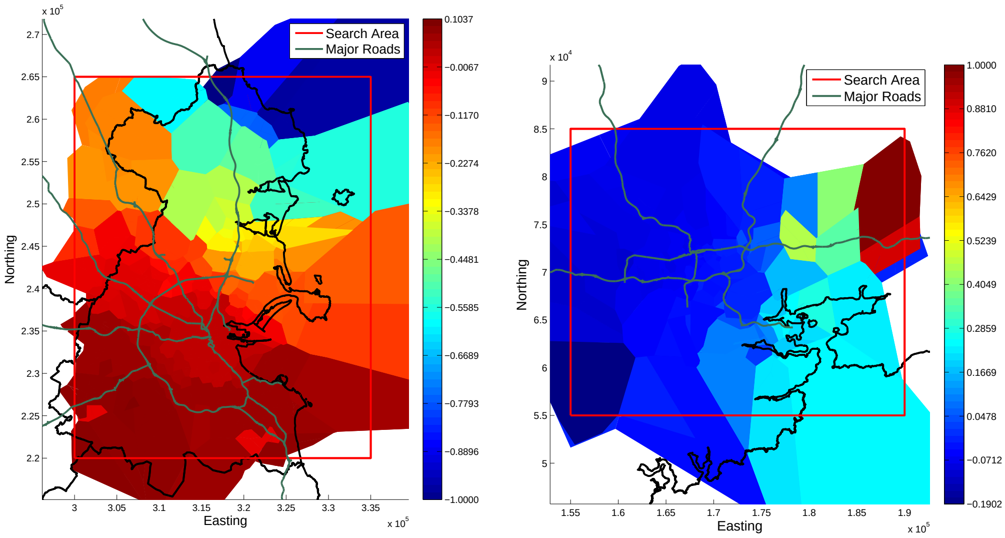


Figure 5.20: Communities identified across the republic of Ireland using the eigenvector of the second largest eigenvalue for the aggregated mobility Markov chain model. Communities are colour coded based on their corresponding eigenvector weight as shown on colour bar. Omitted cell coverage polygons correspond to cells with incorrect location data.

Examining Dublin and Cork cities further, results highlight several sub-communities within each region. As depicted in Figure 5.21, a clear separation exists between sub-communities located in the north and southern regions of county Dublin, while the sub-communities of Cork city tend to flow from east to west. These communities reflect well known social divides and complement results presented by Walsh *et al.* [164], in which the community structure of Dublin was analysed using the flow of mobile communications. In each instance, the mobility Markov chain model was reduced to only capture transition probabilities between individual cell towers inside highlighted search areas.

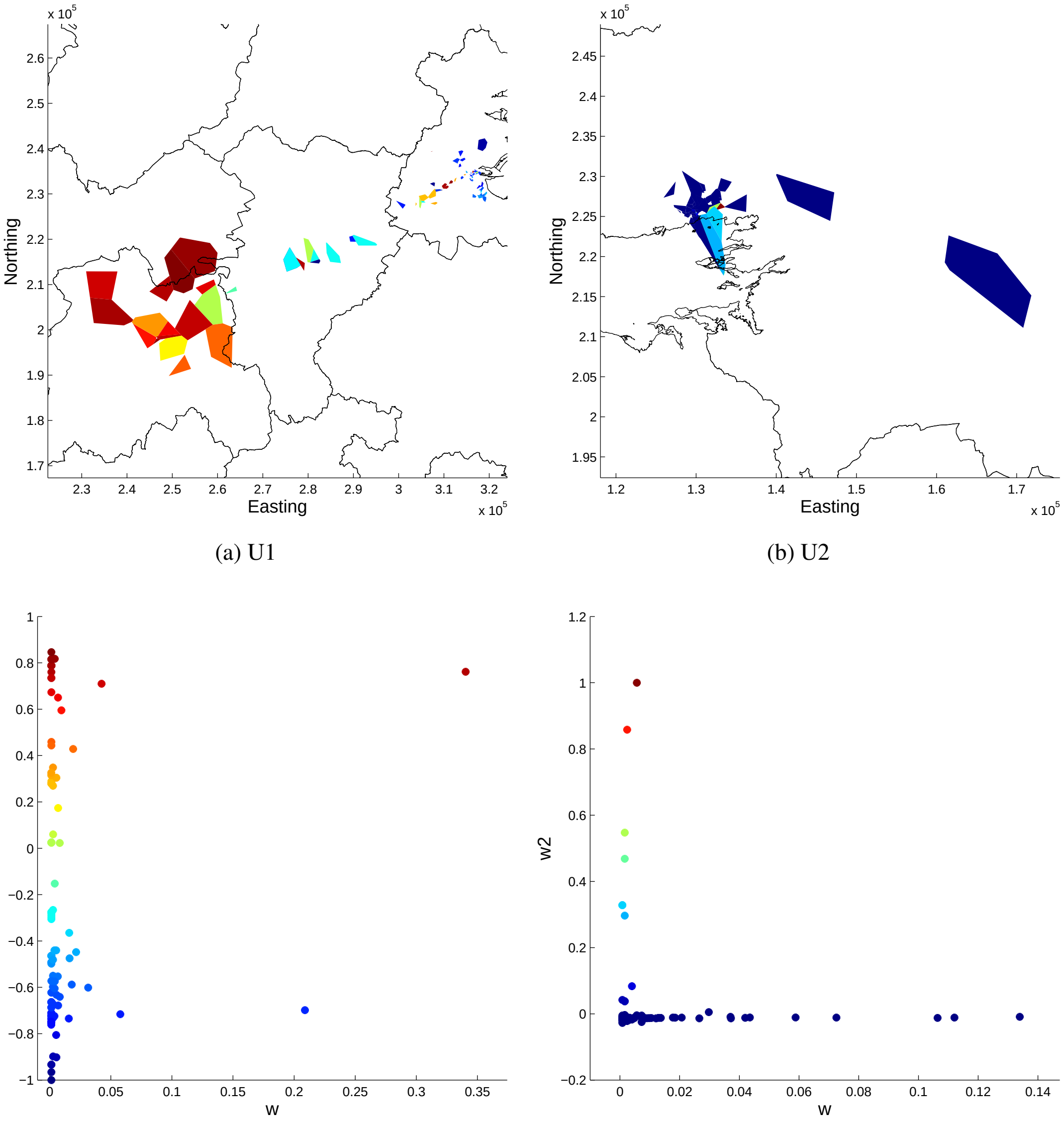


(a) Dublin City

(b) Cork City

Figure 5.21: Visualisation of sub-communities which exist within (a) Dublin city; and (b) Cork city.

Examining the eigenvector of the second largest eigenvalue of a Markov chain mobility model of an individual subscriber, we can investigate if there exists a structure to that subscriber's movement. To illustrate, Figure 5.22 depicts identified communities and corresponding region rank weights for two randomly selected subscribers, U1 and U2. From visual inspection, U1 has two primary communities each centred around cells with large region rank weights. This implies that the subscriber may have a distinct mobility pattern within each community. Alternatively, the community structure identified for U2 indicates that this subscriber's movements are centred around a single community.



(a) U1

(b) U2

(c) U1

(d) U2

Figure 5.22: Identified communities and corresponding region rank weights for two randomly select subscribers, U1 and U2. The communities identified for each subscriber are depicted by (a) and (b), respectively, while the relationship between region rank and second eigenvector weight for each subscriber are depicted by (c) and (d), respectively.

5.5 Discussion

This chapter used call detail records from Meteor, to estimate the regional flows of people across Ireland. Two novel techniques for population estimation based on significant mobile subscriber regions of interest were also introduced. The techniques use the steady state vector of a modified Markov chain mobility model which characterises the mobility of individual subscribers and national aggregated mobility, respectively, as means of identifying the principle location of subscribers, thus providing a proxy for population density.

Results show a high correlation between estimated population density and the national census carried out in 2011. Provided the mobile operator network is servicing a subscriber base proportional to the population both spatially and demographically, the techniques proposed are potential supplements to the procedure of census, which due to the associated costs are infrequently carried out. However, while population fluctuations can be monitored at much finer temporal resolutions, the population metrics which may be captured do not result in the fine grained measurements achieved through a census. Thus, the application of such work may be limited to the estimation of population density and the study of mobility.

Each population density estimation technique discussed has its own advantages and disadvantages. While the estimates derived from the maximum weightings of subscriber Markov chain fixed row vectors are more accurate, the calculation of each individual vector is more computationally intensive compared to the single calculation required from the aggregated vector approach. Moreover, the aggregated approach is totally privacy preserving, as calculations are based on the overall subscriber data instead of individual subscriber regions of interest.

The community structures identified by the eigenvector of the second largest eigenvalue of the aggregated Markov chain mobility model allows regional connectivity patterns to be readily visualised and observed. Similarly, the community structures which exist within an individual subscriber's Markov chain mobility model allows further insight into the subscribers mobility pattern within their environment. However, further research is required to understand the hierarchical structures and temporal dynamics associated with these communities in terms of both methodology and visualisation strategies.

CHAPTER 6

Geographically Located Subscriber Intelligence

Tighter regulation, increasing demand for data services and a fall in the revenues generated from call and SMS traffic means that mobile network operators are beginning to see profit margins fall. In this context, network operators are increasingly focusing their efforts on new revenues generation schemes, lower subscriber churn rate and increasing customer satisfaction. However, such shift in focus has unearthed significant gaps in their knowledge of how subscribers use and perceive the mobile services on offer to them.

This chapter presents initial work into the mining of intelligent geographically located subscriber data. Insights are given into how a mobile network operator may use subscriber generated data to help create new revenue streams and improved network performance.

The chapter is organised as follows. Section 6.1 summarises how information related to subscriber regions of interest may be used by mobile network operators to improve the quality of service delivered to high data users. Section 6.2 then details how the mobility flows related to organised events may be isolated and used to help identify subscriber interests. Section 6.3 describes a methodology which enables targeted geographical marketing applications. Finally, Section 6.4 concludes the chapter with a discussion outlining the future work needed to develop the proposed techniques into suitable commercial applications.

6.1 High Traffic Regions Of Interest

The growing number of smartphones, tablet computers and cellular-network enabled devices means that greater demands are being placed on each mobile network operator to continually deliver a better quality of service to high data driven applications. These services are typically catered for by increasing network capacity in areas where activity spikes are observed. However, little consideration is given to the areas frequented by the subscribers who use the applications.

As previously stated, the steady state vector of a subscriber's modified Markov chain mobility model measures the long term probability of observing that subscriber at a location into the future. By combining the vector weights of subscribers who frequently use large amounts of data, it is possible to get a better understanding of the areas they occupy on a daily basis. This is illustrated in Figure 6.1 showing a map of the combined steady state vector weights for subscribers who were deemed to be high data users.

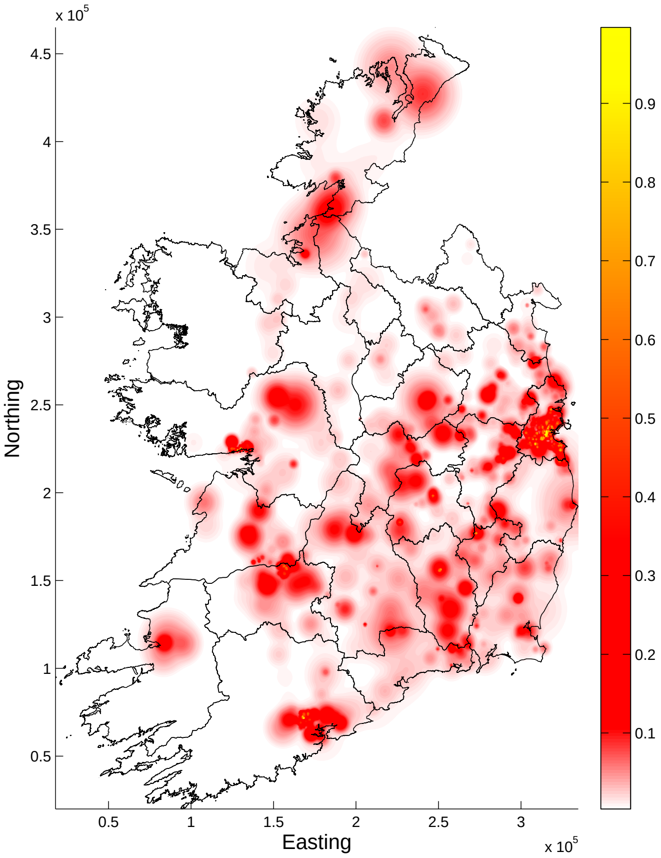


Figure 6.1: Spatial mapping of the combined steady state vector weights from a sample of subscribers deemed to have high data usage behaviour.

Note, to prevent the divulgence of any commercially sensitive information, only a small sample of steady state vector weights were combined and visualised. This map can then be correlated with network coverage maps and customer complaint logs to identify network coverage black spots specific to high demand subscribers.

6.2 Event Mobility

The temporal signatures which relate to large social gatherings can readily be extracted from call detail records (CDR) by observing variations in cell activity and subscriber transition intensity, as such events attract large crowds which can in turn generate lots of above average CDR activities. Illustrating this, Figures 6.2 and 6.3 depict the accumulated temporal activities and the flow of subscribers from cells which service The O2 Dublin amphitheatre from the 01/12/2010 to the 15/01/2011, respectively using a 15-minute temporal bin window. The O2 Dublin is a 14,000 seat amphitheatre located at the Docklands in Dublin City. Here, the flow of subscribers who were stationary within each servicing cell was combined with the flows of subscribers entering and leaving each cell. A summary of the labelled events is given in Table 6.1.

From observing the variations in recorded activities between events, it is clear that the recorded activities are influenced by the social demographics of the subscribers attending each event, as events which cater to juvenile audiences and young adults have much larger activity spikes compared to events which cater to more mature audiences. For example, there are stark differences between the Elton John concert activity spike (H) and the activity spikes corresponding to both JLS concerts (K,L). By identifying the subscribers who attend these particular events mobile network operators can gain valuable insights into their behaviour, which may be exploited for targeted marketing campaigns.

Unlike the event signatures provided by cell activity counts, in many instances the temporal signatures provided by flow intensity are clearly less defined. These differences are likely due to the sampling window applied which restricts observations to within the 15-minute temporal window. This reduces the number of observed transitions at any one instance, as subscribers enter and leave the venue at varying times for different types of events. However, transition intensities can capture many event signatures which cannot be observed using cell tower activity, for example, the flow intensity of subscribers as they enter and leave an area. Also, activity based measurements are subject to network load balancing mechanisms which may result in unexpected bursts in activity as highlighted by event *A* in Figure 6.2. On this date, there was no organised events taking place at The O2 Dublin. Yet, there is a noticeable peak in the activity which is not recorded by the transition intensity based measurements.

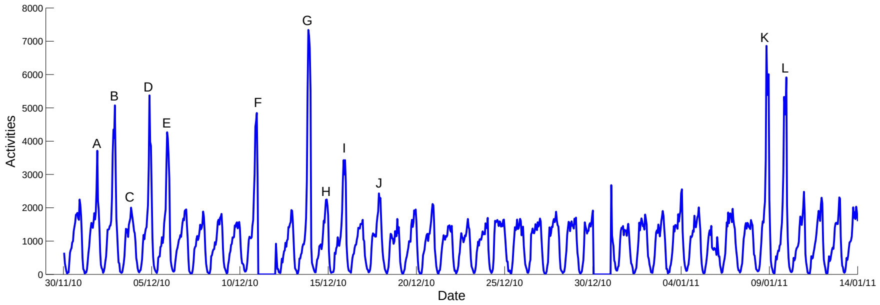


Figure 6.2: The number of CDR activities at cells whose coverage areas polygon services The O2 Dublin, from 01/12/2010 to 15/01/2011 in 15-minute temporal bins.

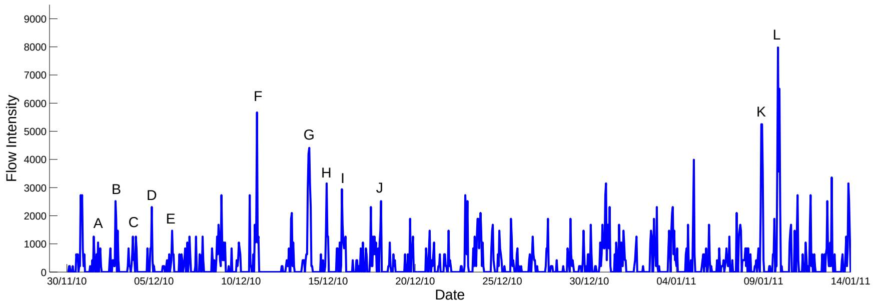
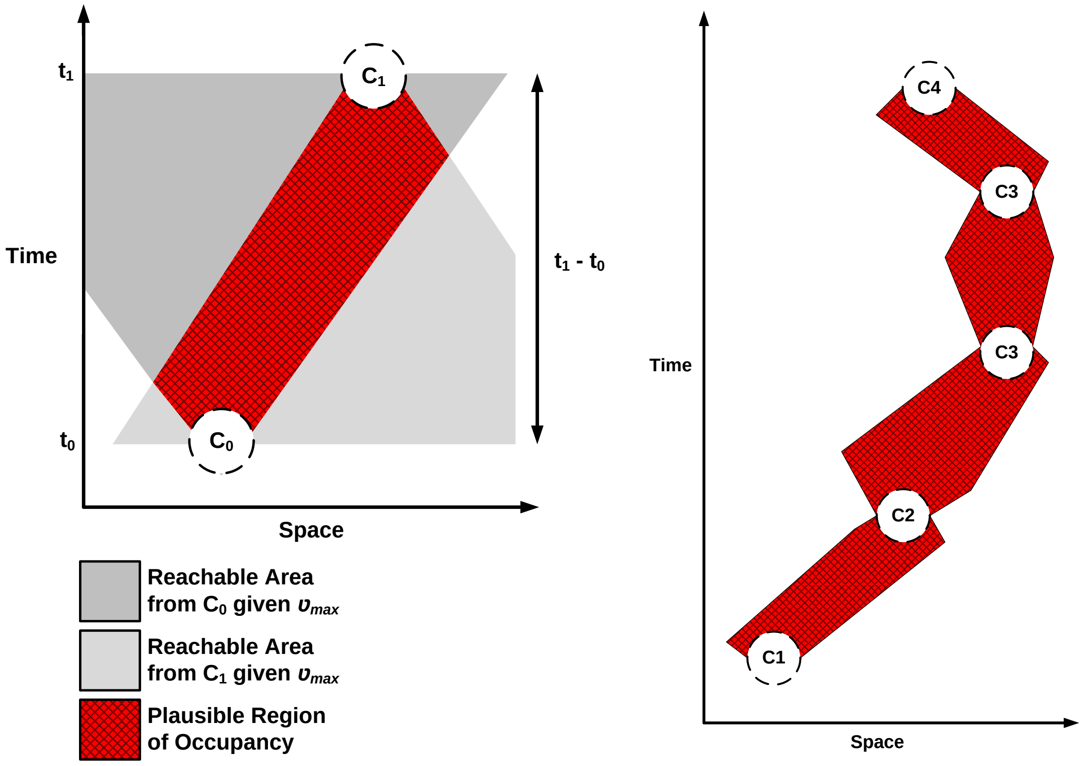


Figure 6.3: The accumulated flow of subscribers around the The O2 Dublin, from the 01/12/2010 to the 15/01/2011 using 15-minute temporal bins.

Table 6.1: Events at The O2 Dublin.

Event	**Date**	**Act**
A	2nd Dec 2010	No Recorded Event
B	3rd Dec 2010	Cheerios ChildLine Concert
C	4th Dec 2010	Horslips
D-E	5th & 6th Dec 2010	Arcade Fire
F	11th Dec 2010	Kings of Leon
G	14th Dec 2010	Deadmau5
H	15th Dec 2010	Elton John
I	16th Dec 2010	Shakira
J	18th Dec 2010	Meat Loaf
K-L	9th & 10th Jan 2010	JLS

A Space Time Prism (STP) [194] is a convenient way to approximate the region of occupancy of an individual between consecutive location samples. Given a maximum velocity v_{max} and a starting cell tower C_0, the region in space which the individual may occupy before being located at the next cell tower C_1 is approximated using the intersection of their STPs. An illustrative example depicting the intersection of STPs is shown in Figure 6.4a. By calculating this region for a series of samples within a trajectory, a Space Time Bead (STB) is formed, as shown by Figure 6.4b. Allowing each location within the plausible region of occupancy to be equally likely, a subscriber may be thought of as being distributed across the region allowing an estimate of temporal spatial density to be taken. Applying this scaling to a subset of subscribers who attend a particular event, their aggregated movement to and from that event may be observed. This information could then be used by event organisers to help with the crowd management of future events. Alternatively, the spatial density of subscribers may be approximated using Gaussian mixture models [216] or kernel weighting functions [195].


t_1
Time
t_0
C_1
C_0
t_1 - t_0
Space
Reachable Area from C_0 given υ_{max}
Reachable Area from C_1 given υ_{max}
Plausible Region of Occupancy
C4
C3
C3
C2
C1
Time
Space

(a) (b)

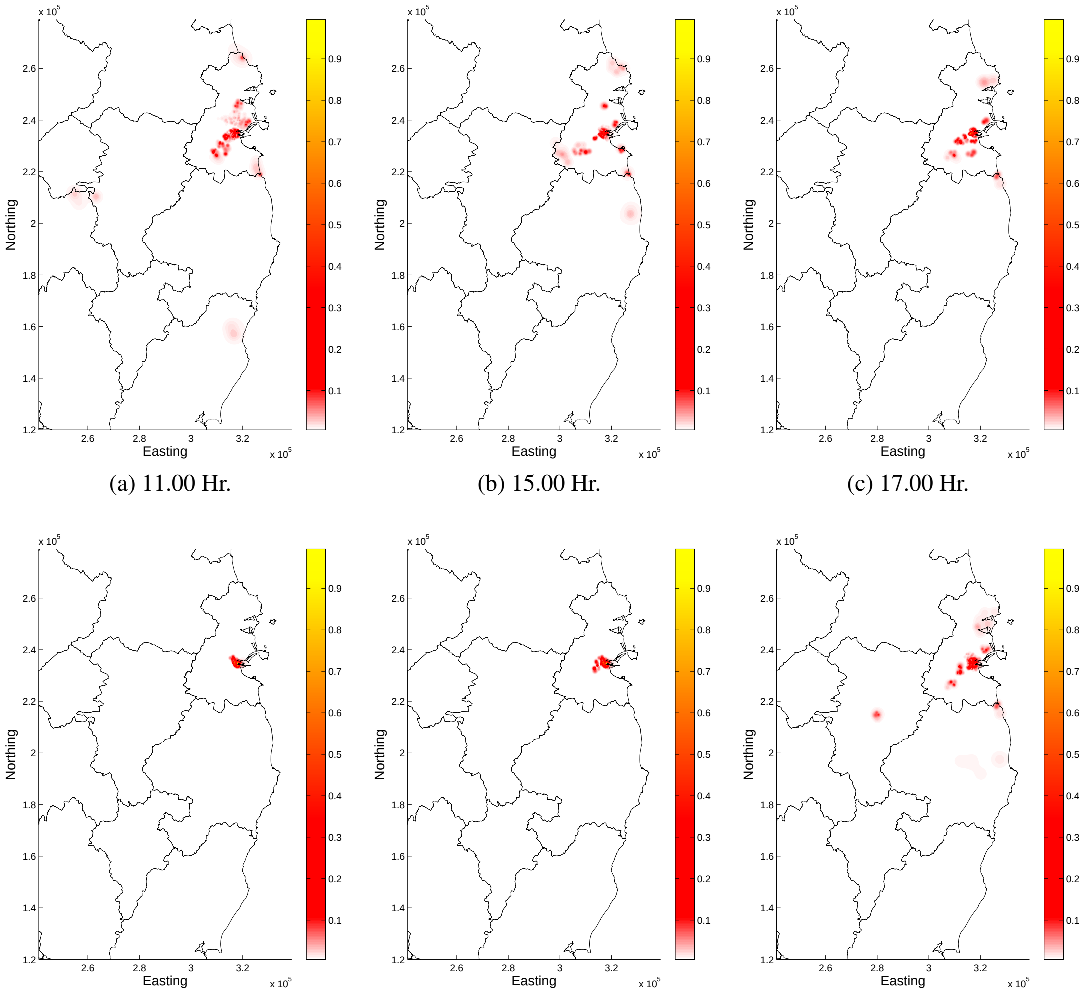
Figure 6.4: Illustration of the plausible region of occupancy estimated using the concept of Space Time Prisms, given a maximum velocity v_{max} and connective locating cell towers C_0 and C_1; (a) Space Time Prisms; (b) Space Time Bead.

As previously discussed, the calculation of velocity using CDR is unreliable due to the uncertainty associated with the arrival time and location estimates. However, a lower bound on velocity $\hat{V}_{ij}$ may be more accurately estimated using observed transition time and the minimum distance between each cell tower's coverage polygon. As such, an estimated value for v_{max} may be obtained by

$$v_{max} = \begin{cases} \hat{V}_{ij} & \text{if } \hat{V}_{ij} \geq \rho \\ \rho & \text{otherwise} \end{cases} \tag{6.1}$$

where ρ is a lower limiting threshold on $\hat{V}_{ij}$. This threshold is introduced to account for situations when $\hat{V}_{ij} \rightarrow 0$ due to instances when there is a large temporal gap between activities that occur between two cells which are in close spatial proximity.

Consider each transition contained within the trajectories of the mobile subscribers who were at or in the vicinity of the O2 Dublin during the Deadmau5 concert, on the 14th December 2010. Using this measure of velocity, it is possible to observe the flux of individuals to and from this event. This movement is shown in Figure 6.5 using KDE estimates of subscriber density given each subscriber's STB. Note, for visual clarity, density normalised such that the maximum density is 1. The depiction of subscriber trajectories in this manner allows the commuting patterns associated with a particular event to be dynamically observed and may be of interest to event marketeers and planners.



(a) 11.00 Hr.

(b) 15.00 Hr.

(c) 17.00 Hr.

(d) 21.00 Hr.

(e) 23.00 Hr.

(f) 2.00 Hr.

Figure 6.5: Kernel density estimates of the changing population density relating to individuals who were at or in the vicinity of the O2 Dublin during the Deadmau5 concert at The O2 amphitheatre Dublin on the 14th December 2010.

6.3 Targeted Geographical Marketing

The integration of trajectory datasets with semantic information is a growing trend in human mobility research [217]. The sources of semantic information varies among different studies. However, recent initiatives such as Dublinked [218] and the growing trend in smart city research [219, 220, 221] means that publicly available datasets are becoming more prevalent. The datasets used in this study consist of information sourced from Dublinked and GeoDirectory [214].

As previously discussed, Section 5.3.1, GeoDirectory is one of the most comprehensive address databases available in the Republic of Ireland. It includes both residential and commercial building locations and typically has limited semantic data related to commercial property use. Dublinked is an Irish data repository set up by Dublin City Council, Dun Laoghaire Rathdown County Council, Fingal County Council and South Dublin County Council in collaboration with NUI Maynooth. Using this resource it is possible to obtain semantic data related to public amenities and services, including shopping centres, recreational areas, health centres, pubs and local eateries.

Due to the uncertainty associated with CDR positional estimates, it is not possible to directly relate CDR samples to particular activities located at individual buildings or amenities. Instead, trajectory samples may be encoded by the semantics within each cell coverage area polygon. This is illustrated in Figure 6.6 through a subscriber CDR trajectory with sample underlying semantic data.

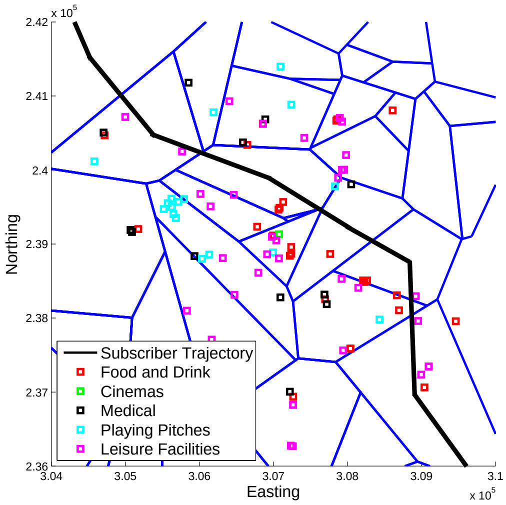
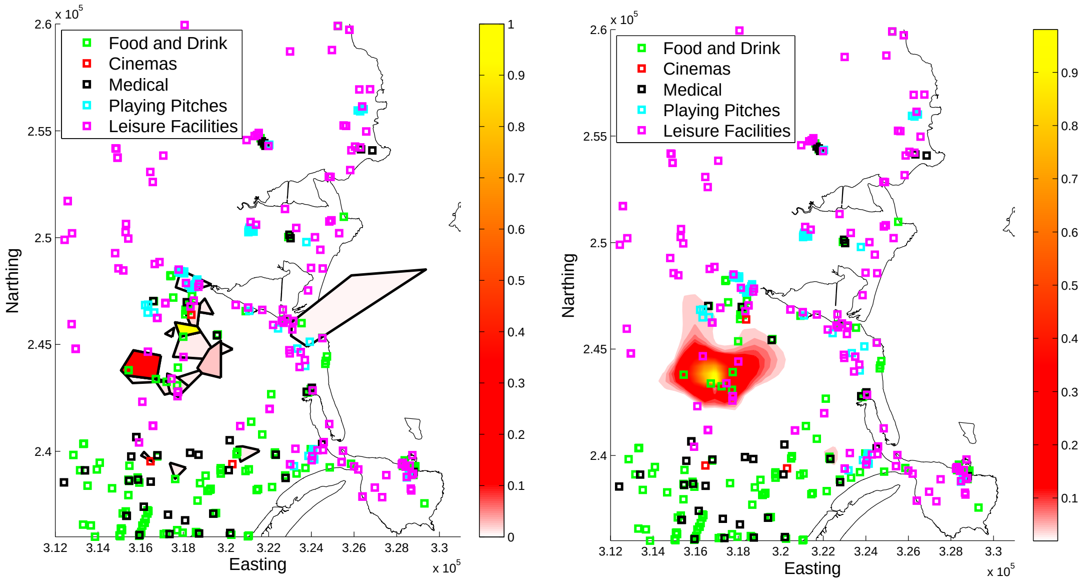


Figure 6.6: Illustration of a subscriber's CDR trajectory with sample semantic data.

It is difficult to determine if subscriber CDR activities are due to that subscriber passing through an area while moving between points of interest, or, as a result of some meaningful activity at that location. As previously demonstrated, the steady state vector of a subscriber's modified Markov chain mobility model is a convenient measure of the importance of each region to an individual subscriber. Generally, people are more likely to use amenities and services which are easily assessable from areas they frequent if they fulfil their specific needs. Therefore, it is reasonable to assume that subscribers steady state vector weights are a reflection of the importance of amenities/services available within a general area to the individual subscriber concerned.

Here, two geographical weighting approaches are applied to measure the relative importance of each amenity to an individual subscriber. The first method calculates the relative importance of an amenity by summing of all steady state vector weights from each cell whose spatial coverage polygon overlaps that amenity. The second approach, ranks each amenity by distributing the steady state vector weights observed at each cell location. This is applied using a weighting function similar to that described in Section 3.3. Illustrating these techniques, Figure 6.7 shows the ranking applied to each amenity/service using the steady vector from a randomly selected subscriber.

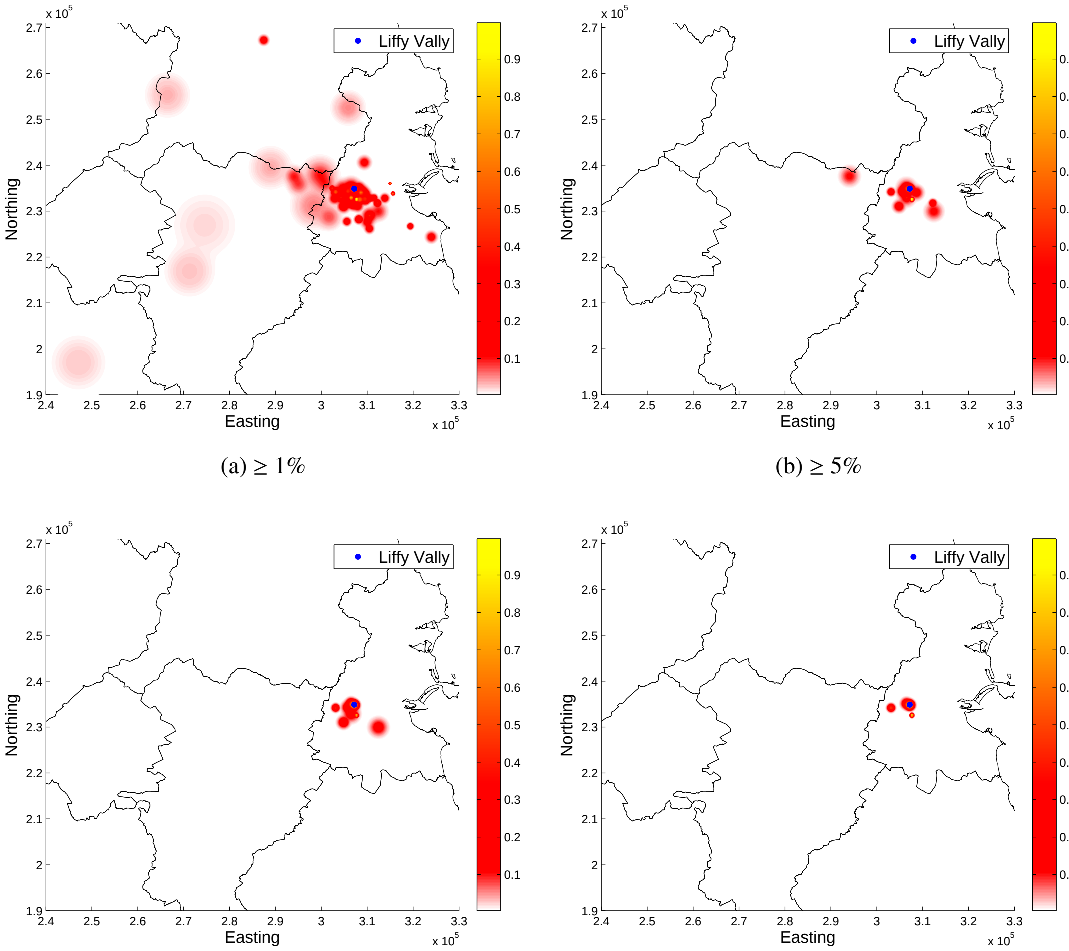


(a) accumulative

(b) ranked

Figure 6.7: The relationship between local amenities and the steady state vector weight, *w*, for a selected subscriber using (a) accumulative cell weighting; and (b) ranked vector weighting.

Estimating the population density of subscribers who have strong links to a particular amenity/service allows the catchment area associated to that particular amenity/service to be determined. The aggregated catchment area data may then be passed on to third parties for targeted marketing purposes, generating new revenue streams for mobile operators. For example, Figure 6.8 depicts the population density of subscribers who have varying levels of attraction to cells covering the Liffey Valley shopping centre, Lucan, Co. Dublin. Note that home location estimates are based on the aforementioned maximum likelihood technique (Section 5.3). Results highlight towns and villages which lie on directly connected roads, with strong links to regions of west Dublin and north east Kildare.



(a) ≥ 1%

(b) ≥ 5%

(c) ≥ 10%

(d) ≥ 20%

Figure 6.8: The catchment area of the Liffey Valley shopping centre, Lucan, Co. Dublin. Estimated of catchment area is based on the population density of subscribers with a steady state vector weight of (a) ≥ 1%; (b) ≥ 5%; (c) ≥ 10% and (d) ≥ 20% relating to cells covering the shopping centre concerned.

6.4 Discussion

This chapter presented initial research into the mining of intelligent geographically located subscriber data using the techniques developed in this thesis. Details are given on how network operators may utilise subscriber mobility steady state vectors to identify coverage black spots. The mapping of combined steady state vectors is a better representation of where subscribers spend time compared to a mapping of home locations, because it also captures a measure of time spent in each region. These insights may help reduce future churn rate, as operators can deliver services to where they are needed.

Methodologies that enable targeted geographical marketing applications are also discussed. This includes a procedure to observe the aggregated flow of subscribers who were identified as being in the vicinity of organised events, and a methodology to geographically weight amenities/services based on subscriber regions of interest.

By incorporating external data sources and better semantic data, some of the ambiguity associated with what places and services subscribers are using can be removed. While more studies are required for more conclusive answers, such data fusion will result in potentially more accurate subscriber segmentation, which may lead to better geographical marketing applications. However, as previously discussed in Section 1.2, fusing CDR with external data sources can make it easier to determine a subscriber's identity. Therefore careful consideration needs to be given to insure subscriber privacy rights are maintained.

CHAPTER 7

Concluding Summary & Future Work

7.1 Concluding Summary

This work contributes to the area of large scale mobility estimation through the use of mobile telephony call detail records (CDR), with enabling methodology for applications such as population estimation, travel route discovery and geographical marketing being detailed. In the development of this methodology, several challenges related to mobile feature extraction, computational complexity, localisation uncertainty and privacy are addressed.

The thesis begins by giving a brief overview of human activity monitoring applications, with a focused discussion on the scalability of each technique. A modern mobile telephony network is then critiqued for its use as a suitable sensing platform for relating human activity patterns, with an accompanying discussion on the various techniques for mobile telephony data procurement and an overview of current research using such data.

Results from an initial investigation into mobile client RSSI collection showed the potential of accumulative RSSI activity as a proxy for mobile device activity over a small geographical area. However, the cost of building a distributed sensor network which could accurately monitor this metric have proven to be prohibitively expensive.

The primary data source used in this body of work was the call detail records (CDR) of

Meteor, a mobile network operator in the Republic of Ireland. Motivated by the minimal costs associated with gathering data from a mobile network operator, there exists opportunities to monitor mobile device activity at urban and national scales, a feat not possible with RSSI collection. Also, as demand for mobile broadband increases, a substantial increase in data rates will result in the need to deploy even smaller cells through urban areas and towns. This will have the effect of reducing location uncertainty and sampling frequency, problems commonly associated with CDR movement data.

The procedures used to extract and visualise a range of features from CDR, cell tower information and subscriber registration data from a cellular phone network was then given. Techniques for data cleansing, cell coverage area modelling and cell clustering were also presented. Along with these contributions, the estimation of the achievable distance a mobile device may travel over time, a novel activity spatial weighting function for spatio-temporal cell activities and a detailed discussion on the time variability of CDR trajectory sampling were provided.

Analysing the movement patterns readily available in CDR is an attractive proposition for transportation engineers and planning authorities. However, investigations of which route an individual travelled along between points of interest are limited due to the localisation and temporal uncertainty associated with CDR samples. To overcome these issues, the novel distance measurements virtual cell path distance (D_{VCP}) and probabilistic cell connectivity distance (D_{PCC}) were developed. These distances enabled a measurement of similarity between CDR journey trajectories and travel paths of interest. Corresponding temporal extensions were also presented, namely modified virtual cell path distance (D_{MVCP}) and modified probabilistic cell connectivity distance (D_{MPCC}).

Evaluated using a test dataset, results showed that both D_{VCP} and D_{PCC} achieved greater accuracy when classifying which route CDR journey trajectories took when compared to traditional trajectory distance measurements. The dataset used was comprised of simulated journey trajectories which were generated using a novel agent based model which simulates CDR journey trajectories between regions of interest. Results also showed that incorporating D_{VCP} or D_{PCC} into the longest common subsequence distance (D_{LCSS}) improves the accuracy of D_{LCSS} with respect to CDR trajectory distance calculations.

Using similar simulated journey trajectories, the performance of D_{MVCP}, D_{MPCC}, D_{LCSS},

D_{LVCP} and D_{LPCC} was compared. Results showed that both D_{MVCP} and D_{MPCC} outperformed D_{LCSS}, D_{LVCP} and D_{LPCC}. However, a limitation of D_{MVCP} and D_{MPCC} is that each distance measurement is sensitive to the noise which may be introduced by outdated cell tower locations. Therefore, when deciding which similarity metric to apply, there is a tradeoff between accuracy and noise tolerance as D_{LVCP} and D_{LPCC} will naturally account for noisy samples.

As a more cost effective alternative to traditional national census, recent studies have demonstrated that the ubiquity of mobile devices may be exploited for population count estimation. However, techniques to date often require the estimation of mobile subscriber home locations and tend to be computationally intensive. As a result, there is a need for techniques which can provide population density measurements which are both computationally efficient and privacy preserving.

To address such concerns, it was shown that the steady state vector of a modified Markov chain mobility model which characterises the aggregated movements of subscribers could be used as an approximation for regional population density. As a matter verification, results showed a high correlation between approximated population density and that measured using a census. Alternatively, the steady state vector of a modified Markov chain mobility model characterising the movement of an individual subscriber was evaluated as a means of quantifying the significance of spatial regions to that individual. Using the region which has the maximum steady state vector weighting as an estimate of the location of his/her home, approximated population density was shown to be highly correlated with that measured using a census.

These findings suggest that provided the mobile operator network is servicing a subscriber base proportional to the population both spatially and demographically, the techniques proposed are potential alternatives to the procedure of census. However, while population fluctuations can be monitored at a much finer temporal resolution, the population metrics which may be captured do not result in the fine grained measurements achievable through a census. Thus, the application of such work may be limited to the estimation of population density and the study of mobility in a regional context.

Initial research into geographically located subscriber insights demonstrated a methodology to segment subscribers based on their perceived relationship to amenities and services. Also investigated was a procedure to identify coverage black spots related to data services and a methodology related to identifying the mobility flows related to organised events. It is envisaged that further research into these applications may help mobile network operators reduce future churn rate, create new revenue streams and improve services delivered to subscribers.

7.2 Future Work

This thesis addresses many of the central issues surrounding the estimation of population density, the identification of which path a subscriber travelled while moving between regions of interest and the mechanisms which enable geographical marketing applications from mobile telephony CDR. However, there are still issues remaining, which, if addressed correctly, could lead to more accurate population estimates, better travel path identification and commercial geographical marketing applications.

There are several features which need to be optimised if population estimation error is to be minimised. These include, among others, the identification of an optimal observation period, optimal spatial resolution for subscriber-based analysis, a more robust home identification procedure and automatic scaling to population demographics.

The observation period dictates the transition weights contained within each Markov chain mobility model. For the national mobility estimation technique, temporal heterogeneity of regional transition flows may result in the steady state vector becoming outdated. As a result, further research is required to determine the optimal observation window. Similarly, if a subscriber moves their home location during the observation period or significantly alters their movement behaviour, their Markov chain mobility model will not accurately capture their updated mobility patterns. As a result, longitudinal studies are needed to ensure that each subscriber's Markov chain mobility model is temporally homogenous, otherwise the explicit removal of outdated positional data may be required.

As previously discussed, typically mobile network topology for 2G and 3G are designed separately. This results in several cells of varying standards covering a single geographical area. Therefore, to capture the true link between subscriber and regional occupancy, cell grouping is required. Hence, to minimise population estimation error the optimal spatial resolution for regions of interest needs to be determined. Likewise, further research into the relationship between residential location and steady state vector weight is also necessary. This work may be supplemented by the validation of results through a survey of mobile subscribers.

The spatial variation of subscriber penetration on a mobile network results in the non-uniform population sampling, both in a regional and national spatial context. To accurately estimate population density, further research is required into how a mobile network operator's subscriber penetration may be scaled, such that uniform population sampling is achieved. In other words, a method of transforming subscriber distribution to population distribution is required for a representative spatial, and presumably temporal, analysis comparable to a census. This task is not trivial as many non-bill paying subscribers do not register correct personal information, resulting in unreliable subscriber demographic data.

If intelligent transportation systems are to benefit from ongoing research into subscriber travel path prediction, algorithms which can avail of real-time transportation data and streamed CDR records may be able to provide in real-time or near real-time the path or mode of transport a subscriber is taking. Similarly, research into the identification of sub-trajectories within subscriber CDR journey trajectories, may identify if a journey consists of multiple paths instead of one single complex path. This information is important as it can be used to identify multiple modes of transport along a single journey.

By incorporating external data sources, and improved semantic data, some of the ambiguity associated with what places and services subscribers are using can be removed. This will result in potentially more accurate subscriber segmentation, and may lead to potentially better commercial geographical marketing applications. However, such data fusion would make it easier to determine a subscriber's identity, thus efforts should be made to ensure that subscriber privacy rights are respected and not infringed.

References

[1] J. Korhonen, *Introduction to 3G Mobile Communications (Second Edition)*. Artech House, 2003.

[2] Population Classified by Area, vol 1, http://www.cso.ie/en/census/census2011reports/census2011populationclassifiedbyareaformerlyvolumeone/, accessed Aug. 2012.

[3] National Travel Survey, http://www.cso.ie/en/media/csoie/releasespublications/documents/transport/2009/nattravel09.pdf, accessed Oct. 2012.

[4] X. Jiang, Y. Qiu, and W. Richard, "U.S. National Household Travel Survey Used to Validate Exposure Estimates by the Quasi-Induced Exposure Technique," *Transportation Research Record: Journal of the Transportation ResearchBoard*, vol. 2237, pp. 152–159, 2011.

[5] R. Hallowell, "The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study," *International Journal of Service Industry Management*, vol. 7, no. 4, pp. 27–42, 1996.

[6] R. Wüstenhagen and M. Bilharz, "Green energy market development in Germany: effective public policy and emerging customer demand," *Energy Policy*, vol. 34, no. 13, pp. 1681–1696, 2006.

[7] J. Friedmann and C. Weaver, *Territory and function: The evolution of regional planning*. Univ of California Press, 1979.

[8] Y. Wu, "The impact of public opinion on board structure changes, director career progression, and CEO turnover: evidence from CalPERS' corporate governance program," *Journal of Corporate Finance*, vol. 10, no. 1, pp. 199 – 227, 2004.

[9] Y.-H. Hwang and D. R. Fesenmaier, "Multidestination pleasure travel patterns: empirical evidence from the American Travel Survey," *Journal of Travel Research*, vol. 42, no. 2, pp. 166–171, 2003.

[10] F. M. Dieleman, M. Dijst, and G. Burghouwt, "Urban form and travel behaviour: micro-level household attributes and residential context," *Urban Studies*, vol. 39, no. 3, pp. 507–527, 2002.

[11] A. Burns and R. F. Bush, "Marketing research," *Globalization*, vol. 1, p. 7, 2000.

[12] L. Kanuk and C. Berenson, "Mail surveys and response rates: A literature review," *Journal of Marketing Research*, pp. 440–453, 1975.

[13] J. R. Hughes, "Nicotine dependence and WHO mental health surveys," *JAMA: the journal of the American Medical Association*, vol. 292, no. 9, pp. 1021–1022, 2004.

[14] L. A. Aday and L. J. Cornelius, *Designing and conducting health surveys: a comprehensive guide*. Jossey-Bass, 2006.

[15] S. Moussavi, S. Chatterji, E. Verdes, A. Tandon, V. Patel, and B. Ustun, "Depression, chronic diseases, and decrements in health: results from the World Health Surveys," *The Lancet*, vol. 370, no. 9590, pp. 851–858, 2007.

[16] C. Cardelino, "Daily variability of motor vehicle emissions derived from traffic counter data," *Journal of the Air & Waste Management Association*, vol. 48, no. 7, pp. 637–645, 1998.

[17] D. Bauer, "Estimating origin-destination-matrices depending on the time of the day from high frequent pedestrian entry and exit counts," *Intelligent Transport Systems, IET*, vol. 6, no. 4, pp. 463–473, 2012.

[18] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 1, pp. 63–71, 2005.

[19] Y. Zhang, D. Yao, T. Qiu, L. Peng, and Y. Zhang, “Pedestrian Safety Analysis in Mixed Traffic Conditions Using Video Data,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 4, pp. 1832–1844, Dec. 2012.

[20] B. Morris and M. Trivedi, “A Survey of Vision-Based Trajectory Learning and Analysis for Surveillance,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 8, pp. 1114–1127, Aug. 2008.

[21] J. Doyle, R. Farrell, S. McLoone, T. McCarthy, M. Tahir, and P. Hung, “Utilising mobile phone RSSI metric for Human Activity Detection,” in *Signals and Systems Conference (ISSC 2009), IET Irish*, June 2009, pp. 1–6.

[22] J. Doyle, R. Farrell, S. McLoone, T. McCarthy, and P. Hung, “Extracting Localised Mobile Activity Patterns from Cumulative Mobile Spectrum RSSI,” in *Proceedings of the China-Ireland Information and Communications TechnologiesConference*, 2009, pp. 75–82.

[23] Z. Sun, G. Bebis, and R. Miller, “On-road vehicle detection: a review,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 5, pp. 694–711, 2006.

[24] J. Hwang, J. Kang, Y. Jang, and H. Kim, “Development of novel algorithm and real-time monitoring ambulatory system using Bluetooth module for fall detection in the elderly,” in *Engineering in Medicine and Biology Society, 2004. IEMBS’04. 26th Annual International Conference of the IEEE*, vol. 1, 2004, pp. 2204–2207.

[25] F. Calabrese and C. Ratti, “Real time rome,” *Networks and Communication studies*, vol. 20, pp. 247–258, 2006.

[26] A. Sevtsuk, S. Huang, F. Calabrese, and C. Ratti, “Mapping the MIT campus in real time using WiFi,” *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*, pp. 326–338, 2008.

[27] N. Andrienko, G. Andrienko, H. Stange, T. Liebig, and D. Hecker, “Visual Analytics for Understanding Spatial Situations from Episodic Movement Data,” *Kunstl Intelligenz*, pp. 1–11, 2012.

[28] D. Brockmann, L. Hufnagel, and T. Geisel, “The scaling laws of human travel,” *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[29] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, “Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system,” *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455 – 466, 2010.

[30] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, “NextPlace: a spatio-temporal prediction framework for pervasive systems,” in *Pervasive Computing*. Springer, 2011, pp. 152–169.

[31] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen, “Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen,” in *Intelligent Transportation Systems, 2009. ITSC’09. 12th International IEEE Conference on*, 2009, pp. 1–6.

[32] R. Ahas, A. Aasa, Ü. Mark, T. Pae, and A. Kull, “Seasonal tourism spaces in Estonia: Case study with mobile positioning data,” *Tourism Management*, vol. 28, no. 3, pp. 898–910, 2007.

[33] R. Ahas, A. Aasa, S. Silm, R. Aunap, H. Kalle, and U. Mark, “Mobile positioning in Space-Time behaviour studies: social positioning method experiments in Estonia,” *Cartography and Geographic Information Science*, vol. 34, no. 4, pp. 259–273, 2007.

[34] A. W. Allaway, R. M. Gooner, D. Berkowitz, and L. Davis, “Deriving and exploring behavior segments within a retail loyalty card program,” *European Journal of Marketing*, vol. 40, no. 11/12, pp. 1317–1339, 2006.

[35] M. T. Capizzi and R. Ferguson, “Loyalty trends for the twenty-first century,” *Journal of Consumer Marketing*, vol. 22, no. 2, pp. 72–80, 2005.

[36] T. M. Mitchell, “Machine learning and data mining,” *Communications of the ACM*, vol. 42, no. 11, pp. 30–36, 1999.

[37] G. Dong, *Sequence data mining*. Springer-Verlag, 2009.

[38] M. J. Berry and G. S. Linoff, *Data mining techniques: for marketing, sales, and customer relationship management*. Wiley Computer Publishing, 2004.

[39] B. OConnor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010, pp. 122–129.

[40] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[41] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer, “Political polarization on twitter,” in *Proc. 5th Intl. Conference on Weblogs and Social Media*, 2011.

[42] A. Sadilek, H. Kautz, and J. P. Bigham, “Finding your friends and following them to where you are,” in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 723–732.

[43] A. Pozdnoukhov and C. Kaiser, “Space-time dynamics of topics in streaming text,” in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. ACM, 2011, pp. 1–8.

[44] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, “A tale of many cities: universal patterns in human urban mobility,” *PloS one*, vol. 7, no. 5, 2012.

[45] G. McArdle, A. Lawlor, E. Furey, and A. Pozdnoukhov, “City-scale traffic simulation from digital footprints,” in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM, 2012, pp. 47–54.

[46] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: user movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[47] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” *Computer Networks*, vol. 38, no. 4, pp. 393 – 422, 2002.

[48] T. Choudhury, S. Consolvo, B. Harrison, J. Hightower, A. LaMarca, L. LeGrand, A. Rahimi, A. Rea, G. Bordello, B. Hemingway *et al.*, "The mobile sensing platform: An embedded activity recognition system," *Pervasive Computing, IEEE*, vol. 7, no. 2, pp. 32–41, 2008.

[49] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "BikeNet: A mobile sensing system for cyclist experience mapping," *ACM Trans. Sen. Netw.*, vol. 6, no. 1, pp. 1–39, Jan. 2010.

[50] A. Milenković, C. Otto, and E. Jovanov, "Wireless sensor networks for personal health monitoring: Issues and an implementation," *Computer communications*, vol. 29, no. 13, pp. 2521–2533, 2006.

[51] Y. Hao and R. Foster, "Wireless body sensor networks for health-monitoring applications," *Physiological measurement*, vol. 29, no. 11, 2008.

[52] D. Raskovic, T. Martin, and E. Jovanov, "Medical monitoring applications for wearable computing," *The Computer Journal*, vol. 47, no. 4, pp. 495–504, 2004.

[53] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *Communications Magazine, IEEE*, vol. 48, no. 9, pp. 140–150, 2010.

[54] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting Space through Digital Signatures," *Pervasive Computing, IEEE*, vol. 9, no. 1, pp. 78–84, 2010.

[55] D. Cox, V. Kindratenko, and D. Pointer, "IntelliBadge TM: towards providing location-aware value-added services at academic conferences," in *UbiComp 2003: Ubiquitous Computing*. Springer, 2003, pp. 264–280.

[56] M. Callaghan, P. Gormley, M. McBride, J. Harkin, and T. McGinnity, "Internal Location Based Services using Wireless Sensor Networks and RFID Technology," *IJCSNS*, vol. 6, no. 4, p. 108, 2006.

[57] D. Kelly, S. McLoone, and T. Dishongh, "A bluetooth-based minimum infrastructure home localisation system," in *Wireless Communication Systems. 2008. ISWCS'08. IEEE International Symposium on*, 2008, pp. 638–642.

[58] G. Andrienko, N. Andrienko, and M. Heurich, “An event-based conceptual model for context-aware movement analysis,” *International Journal of Geographical Information Science*, vol. 25, no. 9, pp. 1347–1370, 2011.

[59] Y. Zheng, L. Liu, L. Wang, and X. Xie, “Learning transportation mode from raw gps data for geographic applications on the web,” in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 247–256.

[60] S. Feng and C. L. Law, “Assisted GPS and its impact on navigation in intelligent transportation systems,” in *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*, 2002, pp. 926–931.

[61] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, “Learning and inferring transportation routines,” *Artificial Intelligence*, vol. 171, pp. 311 – 331, 2007.

[62] E. Lee, M. Y. Hu, and R. S. Toh, “Are consumer survey results distorted? Systematic impact of behavioral frequency and duration on survey response errors,” *Journal of Marketing Research*, pp. 125–133, 2000.

[63] G. Mao, B. Fidan, and B. Anderson, “Wireless sensor network localization techniques,” *Computer Networks*, vol. 51, pp. 2529–2553, 2007.

[64] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, “Are call detail records biased for sampling human mobility?” *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 16, no. 3, pp. 33–44, Dec. 2012.

[65] A. Mishra, *Fundamentals of Cellular Network Planning and Optimisation: 2g/2.5g/3g... Evolution to 4g*. John Wiley & Sons, 2004.

[66] B. Rao and L. Minakakis, “Evolution of mobile location-based services,” *Commun. ACM*, vol. 46, no. 12, pp. 61–65, Dec. 2003.

[67] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, “Cellular census: Explorations in urban data collection,” *IEEE Pervasive Computing*, vol. 6, no. 3, pp. 30–38, 2007.

[68] R. Ahas, S. Silm, O. Jaumlrv, E. Saluveer, and M. Tiru, "Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3 – 27, 2010.

[69] N. Eagle, J. A. Quinn, and A. Clauset, "Methodologies for Continuous Cellular Tower Data Analysis," in *Pervasive*, 2009, pp. 342–353.

[70] N. Eagle, A. Clauset, and J. Quinn, "Location Segmentation, Inference and Prediction for Anticipatory Computing," in *Proceedings of AAAI Spring Symposium on Technosocial Predictive Analytics, Stanford, CA*, 2009.

[71] J. Steenbruggen, M. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities," *GeoJournal*, pp. 1–21, 2011.

[72] A. Westin, "Privacy and freedom," *Washington and Lee Law Review*, vol. 25, no. 1, p. 166, 1968.

[73] F. Bignami, "Privacy and Law Enforcement in the European Union: The Data Retention Directive," *Chicago Journal of International Law, Spring*, 2007.

[74] Charter of Fundemental Rights of the European Union (2000/ C364/01).

[75] European-Commission, "EU Directive 1995/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data." *Official Journal of the European Communities L 281*, 1995.

[76] ——, "DIRECTIVE 2002/58/EC concerning the processing of personal data and the protection of privacy in the electronic communications sector," *Official Journal of the European Communities L 281*, 2002.

[77] ——, "EU Directive 2006/24/EC on the retention of data generated or processed in connection with th provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC," *Official Journal of the European Communities L 105*, 2006.

[78] S. Gambs, M.-O. Killijian, and M. N. n. del Prado Cortez, “Show me how you move and I will tell you who you are,” in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. ACM, 2010, pp. 34–41.

[79] S. Gambs, O. Heen, and C. Potin, “A comparative privacy analysis of geosocial networks,” in *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*. ACM, 2011, pp. 33–40.

[80] M. Barbaro and T. Zeller, “A Face Is Exposed for AOL Searcher No. 4417749,” *New York Times*, Aug. 2006.

[81] L. Sweeney, “k-anonymity: A model for protecting privacy,” *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[82] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Incognito: efficient full-domain K-anonymity,” in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 49–60.

[83] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Mondrian Multidimensional K-Anonymity,” in *Data Engineering, 2006. ICDE ’06. Proceedings of the 22nd International Conference on*, April 2006, p. 25.

[84] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[85] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,” in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, April 2007, pp. 106 –115.

[86] M. González, C. Hidalgo, and A. Barabási, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[87] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, “Limits of Predictability in Human Mobility,” *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[88] R. Ahas, U. Mark, O. Jarv, and M. Nuga, "Mobile positioning in sustainability studies: The social positioning method in studying commuter's activity spaces in Tallinn," *WIT Transactions on Ecology and the Environment*, vol. 93, pp. 127–135, 2006.

[89] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sept. 2010, pp. 312 – 317.

[90] J. Onnela, J. Saramaki, J. Hyvonen, G. Szabó, D. Lazer, K. Kaski, J. Kertesz, and A. Barabasi, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, p. 7332, 2007.

[91] J. Reades, F. Calabrese, and C. Ratti, "Eigenplaces: analysing cities using the space-time structure of the mobile phone network," *Environment and Planning B: Planning and Design*, vol. 36, no. 5, pp. 824–836, 2009.

[92] R. Caceres, J. Rowland, C. Small, and S. Urbanek, "Exploring the Use of Urban Greenspace through Cellular Network Activity," *In Proc. of 2nd Workshop on Pervasive Urban Applications (PURBA)*, June 2012.

[93] C. Theodore, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[94] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2010.

[95] M. Olsson, S. Sultana, S. Rommer, L. Frid, and C. Mulligan, *SAE and the Evolved Packet Core: Driving the Mobile Broadband Revolution.* Academic Press, 2009.

[96] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *Communications Surveys Tutorials, IEEE*, vol. 3, no. 2, pp. 10–31, 2000.

[97] G. Sun, J. Chen, W. Guo, and K. Liu, "Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs," *Signal Processing Magazine, IEEE*, vol. 22, no. 4, pp. 12–23, 2005.

[98] Y. Zhao, "Standardization of mobile phone positioning for 3g systems," *Communications Magazine, IEEE*, vol. 40, no. 7, pp. 108 –116, jul 2002.

[99] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of Wireless Indoor Positioning Techniques and Systems," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1067 –1080, Nov. 2007.

[100] K. Kaemarungsi and P. Krishnamurthy, "Modeling of indoor positioning systems based on location fingerprinting," in *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2. IEEE, 2004, pp. 1012–1022.

[101] K. Pahlavan, X. Li, and J. Makela, "Indoor geolocation science and technology," *Communications Magazine, IEEE*, vol. 40, no. 2, pp. 112–118, 2002.

[102] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, vol. 34, no. 8, pp. 57–66, Aug. 2001.

[103] I. Jami, M. Ali, and R. Ormondroyd, "Comparison of methods of locating and tracking cellular mobiles," in *Novel Methods of Location and Tracking of Cellular Mobiles and Their System Applications (Ref. No. 1999/046), IEE Colloquium on*, 1999, pp. 1–6.

[104] A. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based wireless location: challenges faced in developing techniques for accurate wireless location information," *Signal Processing Magazine, IEEE*, vol. 22, no. 4, pp. 24 – 40, July 2005.

[105] C. Morelli, M. Nicoli, V. Rampa, and U. Spagnolini, "Hidden Markov Models for Radio Localization in Mixed LOS/NLOS Conditions," *Signal Processing, IEEE Transactions on*, vol. 55, no. 4, pp. 1525 –1542, Apr. 2007.

[106] C.-D. Wann and M.-H. Lin, "Data fusion methods for accuracy improvement in wireless location systems," in *Wireless Communications and Networking Conference, 2004. WCNC. 2004 IEEE*, vol. 1, Mar. 2004, pp. 471–476.

[107] A. Urruela and J. Riba, "Efficient mobile location from time measurements with unknown variances in dynamic scenarios," in *Signal Processing Advances in Wireless Communications, 2004 IEEE 5th Workshop on*, July 2004, pp. 571 – 575.

[108] T. Roos, P. Myllymaki, and H. Tirri, “A statistical modeling approach to location estimation,” *Mobile Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 59 – 69, 2002.

[109] F. Gustafsson and F. Gunnarsson, “Mobile positioning using wireless networks: possibilities and fundamental limitations based on available wireless network measurements,” *Signal Processing Magazine, IEEE*, vol. 22, no. 4, pp. 41–53, 2005.

[110] B. Mark and Z. Zaidi, “Robust mobility tracking for cellular networks,” in *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 1, 2002, pp. 445 –449.

[111] V. Seshadri, G. Zaruba, and M. Huber, “A Bayesian sampling approach to in-door localization of wireless devices using received signal strength indication,” in *Pervasive Computing and Communications, 2005. PerCom 2005. Third IEEE International Conference on*, Mar. 2005, pp. 75–84.

[112] C.-D. Wann, Y.-M. Chen, and M.-S. Lee, “Mobile location tracking with NLOS error mitigation,” in *Global Telecommunications Conference, 2002. GLOBECOM ’02. IEEE*, vol. 2, Nov. 2002, pp. 1688–1692.

[113] M. Najar and J. Vidal, “Kalman tracking for mobile location in NLOS situations,” in *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, vol. 3, Sept. 2003, pp. 2203–2207.

[114] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, “Data for Development: the D4D Challenge on Mobile Phone Data,” *arXiv preprint arXiv:1210.0137*, 2012.

[115] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen, “The mobile data challenge: Big data for mobile computing research,” in *Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing*, 2012, pp. 1–8.

[116] D. Kotz and T. Henderson, “CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth,” *Pervasive Computing, IEEE*, vol. 4, no. 4, pp. 12–14, 2005.

[117] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, “Mobile Landscapes: using location data from cell phones for urban analysis,” *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, vol. 33, no. 5, p. 727, 2006.

[118] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer, “Mobile landscapes: Graz in real time,” *Location based services and telecartography*, pp. 433–444, 2007.

[119] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, “Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 141 –151, Mar. 2011.

[120] T. Horanont and R. Shibasaki, “An Implementation of Mobile Sensing for Large-Scale Urban Monitoring,” *Proc. of Urbansense08*, 2008.

[121] R. Ahas, A. Aasa, S. Silm, and M. Tiru, “Daily rhythms of suburban commuter movements in the Tallinn metropolitan area: Case study with mobile positioning data,” *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 1, pp. 45–54, 2010.

[122] ——, “Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia,” *Information and communication technologies in tourism 2007*, pp. 119–128, 2007.

[123] R. Ahas, A. Aasa, A. Roose, Ü. Mark, and S. Silm, “Evaluating passive mobile positioning data for tourism surveys: An Estonian case study,” *Tourism Management*, vol. 29, no. 3, pp. 469–486, 2008.

[124] A. Kuusik, R. Ahas, and M. Tiru, “The ability of turism events to generate destination loyalty towards the country: an Estonian case study.” *Maeltsamees, S., Reiljan, J. (Eds),Discussions of Estonian Economic Policy XVIII, Berliner Wissenchafts-Verlag, Berlin*, pp. 140–55, 2010.

[125] A. Kuusik, M. Tiru, and R. Ahas, “Innovation in destination marketing: The use of passive mobile positioning for the segmentation of repeat visitors in Estonia,” *Baltic Journal of Management*, vol. 6, 2011.

[126] S. Silm and R. Ahas, “The seasonal variability of population in Estonian municipalities,” *Environment and planning. A*, vol. 42, no. 10, pp. 2527–2546, 2010.

[127] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, “Estimating Origin-Destination Flows Using Mobile Phone Location Data,” *Pervasive Computing, IEEE*, vol. 10, no. 4, pp. 36 –44, Apr. 2011.

[128] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, “Identifying important places in peoples lives from cellular network data,” *Pervasive Computing*, pp. 133–151, 2011.

[129] D. Kelly, J. Doyle, and R. Farrell, “Analysing Ireland’s Social and Transport Networks using Sparse Cellular Network Data,” in *IET Irish Signals and Systems Conference*, 2011.

[130] G. Andrienko, N. Andrienko, M. Mladenov, M. Mock, and C. Plitz, “Discovering bits of place histories from people’s activity traces,” in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, Oct. 2010, pp. 59 –66.

[131] G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C. Politz, and T. Schreck, “A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage,” *Journal of Location Based Services*, vol. 4, no. 3-4, pp. 200–221, 2010.

[132] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, “A Tale of One City: Using Cellular Network Data for Urban Planning,” *Pervasive Computing, IEEE*, vol. 10, no. 4, pp. 18 –26, April 2011.

[133] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky, “A tale of two cities,” in *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*. ACM, Feb. 2010, pp. 19–24.

[134] M. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Mandnez, “Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics,” in *IEEE Second International Conference on Social Computing (SocialCom)*, Aug. 2010, pp. 241 –248.

[135] R. Becker, R. Cáceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, “Clustering Anonymized Mobile Call Detail Records to Find Usage Groups,” *1st Workshop on Pervasive Urban Applications (PURBA)*, June 2011.

[136] Q. Lin and Y. Wan, “Mobile Customer Clustering Based on Call Detail Records for Marketing Campaigns,” in *Management and Service Science, 2009. MASS ’09. International Conference on*, Sep. 2009, pp. 1–4.

[137] C. Song, T. Koren, P. Wang, and A. Barabási, “Modelling the scaling properties of human mobility,” *Nature Physics*, vol. 6, pp. 818–823, 2010.

[138] J. Park, D. S. Lee, and M. C. González, “The eigenmode analysis of human motion,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, no. 11, pp. 11 021–11 036, Nov. 2010.

[139] T. Couronne, A.-M. Olteanu, and Z. Smoreda, “Urban Mobility: Velocity and Uncertainty in Mobile Phone Data,” in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, Oct. 2011, pp. 1425 –1430.

[140] C. Kang, S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, and X. Ma, “Analyzing and geo-visualizing individual human mobility patterns using mobile call records,” in *Geoinformatics, 2010 18th International Conference on*, Jun. 2010, pp. 1 –7.

[141] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, “Human Mobility Modeling at Metropolitan Scales,” *10th ACM International Conference on Mobile Systems, Applications and Services (MobiSys)*, Jun. 2012.

[142] M. R. Vieira, E. Fr andas Mart andnez, P. Bakalov, V. Fr andas Mart andnez, and V. J. Tsotras, “Querying spatio-temporal patterns in mobile phone-call databases,” in *Mobile*

Data Management (MDM), 2010 Eleventh International Conference on, may 2010, pp. 239 –248.

[143] X. Lu, L. Bengtsson, and P. Holme, “Predictability of population displacement after the 2010 Haiti earthquake,” *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 576–11 581, 2012.

[144] S. Phithakkitnukoon and C. Ratti, “Inferring Asymmetry of Inhabitant Flow using Call Detail Records,” *Journal of Advances in Information Technology*, vol. 2, no. 2, 2011.

[145] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, “Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data,” in *Human Behavior Understanding*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 6219, pp. 14–25.

[146] Y. Yim, “The State of Cellular Probes,” Institute of Transportation Studies, UC Berkeley, Institute of Transportation Studies, Research Reports, Working Papers, Proceedings, 2003.

[147] G. Rose, “Mobile Phones as Traffic Probes: Practices, Prospects and Issues,” *Transport Reviews*, vol. 26, no. 3, pp. 275–291, 2006.

[148] J. Doyle, P. Hung, D. Kelly, S. McLoone, and R. Farrell, “Utilising Mobile Phone Billing Records for Travel Mode Discovery,” in *IET Irish Signals and Systems Conference*, 2011.

[149] N. Caceres, J. Wideberg, and F. Benitez, “Review of traffic data estimations extracted from cellular networks,” *Intelligent Transport Systems, IET*, vol. 2, no. 3, pp. 179 –192, Sept. 2008.

[150] N. Caceres, L. Romero, F. Benitez, and J. del Castillo, “Traffic Flow Estimation Models Using Cellular Phone Data,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 3, pp. 1430–1441, 2012.

[151] H. Bar-Gera, “Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel,” *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380 – 391, 2007.

[152] N. Caceres, J. Wideberg, and F. Benitez, “Deriving origin destination data from a mobile phone network,” *Intelligent Transport Systems, IET*, vol. 1, no. 1, pp. 15 –26, March 2007.

[153] J. White and I. Wells, “Extracting origin destination information from mobile phone data,” in *Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486)*, 2002, pp. 30 – 34.

[154] W. Huayong, F. Calabrese, G. Di Lorenzo, and C. Ratti, “Transportation mode inference from anonymized and aggregated mobile phone call detail records,” in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sep. 2010, pp. 318 –323.

[155] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, “Transportation mode inference from anonymized and aggregated mobile phone call detail records,” in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sep. 2010, pp. 318 –323.

[156] N. Eagle, Y.-A. de Montjoye, and L. Bettencourt, “Community Computing: Comparisons between Rural and Urban Societies Using Mobile Phone Data,” in *Computational Science and Engineering, 2009. CSE ’09. International Conference on*, vol. 4, Aug. 2009, pp. 144 –150.

[157] E. Frias Martinez, G. Williamson, and V. Frias-Martinez, “An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information,” in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, Oct. 2011, pp. 57 –64.

[158] J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski, A. Barabási, and J. Kertész, “Analysis of a large-scale weighted network of one-to-one human communication,” *New Journal of Physics*, vol. 9, p. 179, 2007.

[159] M. Kamola, E. Niewiadomska-Szynkiewicz, and B. Piech, “Reconstruction of a social network graph from incomplete call detail records,” in *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, Oct. 2011, pp. 136–140.

[160] A. Nanavati, R. Singh, D. Chakraborty, K. Dasgupta, S. Mukherjea, G. Das, S. Gurumurthy, and A. Joshi, “Analyzing the Structure and Evolution of Massive Telecom Graphs,” *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 5, pp. 703–718, May 2008.

[161] R. Lambiotte, V. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, “Geographical dispersal of mobile communication networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

[162] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel, “Urban gravity: a model for inter-city telecommunication flows,” *Journal of Statistical Mechanics: Theory and Experiment*, 2009.

[163] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, “Redrawing the Map of Great Britain from a Network of Human Interactions,” *PLoS ONE*, vol. 5, no. 12, 12 2010.

[164] F. Walsh and A. Pozdnoukhov, “Spatial structure and dynamics of urban communities,” *Pervasive Urban Applications workshop at PERVASIVE 2011*, 2011.

[165] R. Ahas and U. Mark, “Location based services a new challenges for planning and public administration,” *Futures*, vol. 37, no. 6, pp. 547 – 561, 2005.

[166] N. Eagle and A. (Sandy) Pentland, “Reality mining: sensing complex social systems,” *Personal Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, Mar. 2006.

[167] N. Eagle, A. S. Pentland, and D. Lazer, “Mobile phone data for inferring social network structure,” in *Social computing, behavioral modeling, and prediction*. Springer, 2008, pp. 79–88.

[168] H. Zhuang, J. Tang, W. Tang, T. Lou, A. Chin, and X. Wang, “Actively learning to infer social ties,” *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 270–297, 2012.

[169] M. De Domenico, A. Lima, and M. Musolesi, “Interdependence and predictability of human mobility and social interactions,” in *Proceedings of the Nokia Mobile Data Challenge 2012 Workshop, Newcastle, United Kingdom.*, 2012.

[170] G. Chittaranjan, J. Blom, and D. Gatica-Perez, “Mining large-scale smartphone data for personality studies,” *Personal Ubiquitous Comput.*, vol. 17, no. 3, pp. 433–450, Mar. 2013.

[171] Path Intelligence, http://www.pathintelligence.com/website-prodnserv.htm, accessed Aug. 2011.

[172] The Irish Grid A Description of the Coordinate Reference System Used in Ireland, http://www.osi.ie/Services/GPS-Services/Reference-Information/Irish-Grid-Reference-System.aspx, accessed Sept. 2011.

[173] A. Okabe, B. Boots, K. Sugihara, and S. Chiu, *Spatial tessellations: concepts and applications of Voronoi diagrams*. Wiley & Sons Chichester., 1992.

[174] Ordinance Survey Ireland, http://www.osi.ie/Home.aspx.

[175] Average Travel Speeds in Northern Ireland, http://www.dhsspsni.gov.uk/paperspeeds.pdf, accessed Oct. 2012.

[176] M. Kraak, “The space-time cube revisited from a geovisualization perspective,” in *Proc. 21st International Cartographic Conference*, 2003, pp. 1988–1996.

[177] H. Varian, “Bootstrap tutorial,” *Mathematica Journal*, vol. 9, no. 4, pp. 768–775, 2005.

[178] B. Welch, “The generalization ofstudent’s’ problem when several different population variances are involved,” *Biometrika*, pp. 28–35, 1947.

[179] R. Xu and I. Wunsch, D., “Survey of clustering algorithms,” *Neural Networks, IEEE Transactions on*, vol. 16, no. 3, pp. 645 –678, May 2005.

[180] G. Andrienko and N. Andrienko, “A general framework for using aggregation in visual exploration of movement data,” *Cartographic Journal, The*, vol. 47, no. 1, pp. 22–40, 2002.

[181] N. Adrienko and G. Adrienko, “Spatial Generalization and Aggregation of Massive Movement Data,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 2, pp. 205–219, Feb. 2011.

[182] K. Buchin, B. Speckmann, and K. Verbeek, “Flow Map Layout via Spiral Trees,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2536 –2544, Dec. 2011.

[183] ORA, vol 1, http://www.casos.cs.cmu.edu/projects/ora/, accessed Aug . 2012.

[184] P. Mutton and J. Golbeck, “Visualization of semantic metadata and ontologies,” in *Information Visualization, 2003. IV 2003. Proceedings. Seventh International Conference on*, July 2003, pp. 300–305.

[185] J. Hajek, J. Billing, and D. Swan, “Forecasting Traffic Loads for Mechanistic-Empirical Pavement Design,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2256, pp. 151–158, 2011.

[186] R. A. Johnston and T. de la Barra, “Comprehensive regional modeling for long-range planning: linking integrated urban models and geographic information systems,” *Transportation Research Part A: Policy and Practice*, vol. 34, no. 2, pp. 125 – 136, 2000.

[187] D. Pickrell and P. Schime, “Growth in motor vehicle ownership and use: Evidence from the nationwide personal transportation survey,” *Journal of Transportation and Statistics*, vol. 2, 1999.

[188] N. Ferdous, R. Pendyala, C. Bhat, and K. Konduri, “Modeling the Influence of Family, Social Context, and Spatial Proximity on Use of Nonmotorized Transport Mode,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2230, pp. 111–120, 2011.

[189] M. Fontaine, A. Yakkala, and B. Smith, “Probe Sampling Strategies for Traffic Monitoring Systems Based on Wireless Location Technology,” Tech. Rep., 2007.

[190] A. Narayanan and V. Shmatikov, “De-anonymizing Social Networks,” in *30th IEEE Symposium on Security and Privacy*, May 2009, pp. 173 –187.

[191] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces," in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 2010, pp. 185–196.

[192] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless communications and mobile computing*, vol. 2, no. 5, pp. 483–502, 2002.

[193] F. Bai and A. Helmy, "A survey of mobility models," *Wireless Adhoc Networks. University of Southern California, USA*, vol. 206, 2004.

[194] H. J. MILLER, "Modelling accessibility using space-time prism concepts within geographical information systems," *International journal of geographical information systems*, vol. 5, no. 3, pp. 287–301, 1991.

[195] U. Demšar and K. Virrantaus, "Space-time density of trajectories: exploring spatio-temporal patterns in movement data," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1527–1542, 2010.

[196] OpenCellID, http://opencellid.org, accessed Sept. 2011.

[197] A. Hurford, "GPS Measurement Error Gives Rise to Spurious 180 Turning Angles and Strong Directional Biases in Animal Movement Data," *PLoS ONE*, vol. 4, no. 5, May 2009.

[198] J. Lou, Q. Liu, T. Tan, and W. Hu, "Semantic interpretation of object activities in a surveillance system," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 3, 2002, pp. 777–780.

[199] M.-P. Dubuisson and A. Jain, "A modified Hausdorff distance for object matching," in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, Oct. 1994, pp. 566–568.

[200] E. J. Keogh and M. J. Pazzani, "Scaling up Dynamic Time Warping for Datamining Applications," in *In Proc. 6th Int. Conf. on Knowledge Discovery and Data Mining*, 2000, pp. 285–289.

[201] B. T. Morris and M. M. Trivedi, “Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.

[202] D. S. Hirschberg, “Algorithms for the longest common subsequence problem,” *J. ACM*, vol. 24, no. 4, pp. 664–675, Oct. 1977.

[203] S. Dodge, P. Laube, and R. Weibel, “Movement similarity assessment using symbolic representation of trajectories,” *International Journal of Geographical Information Science*, pp. 1–26, 2012.

[204] Z. Zhang, K. Huang, and T. Tan, “Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 3, 2006, pp. 1135 –1138.

[205] J. Yen, “Finding the k shortest loopless paths in a network,” *management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[206] M. Buchin, A. Driemel, M. van Kreveld, and V. Sacristan, “An Algorithmic Framework for Segmenting Trajectories based on Spatio-Temporal Criteria,” in *Proc. 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS)*. ACM, 2010, pp. 202–211.

[207] E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[208] R. Ocañ-Riola, “Non-homogeneous Markov Processes for Biomedical Data Analysis,” *Biometrical journal*, vol. 47, no. 3, pp. 369–376, 2005.

[209] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257 –286, Feb. 1989.

[210] M. Ostendorf, V. Digalakis, and O. Kimball, “From hmm’s to segment models: a unified view of stochastic modeling for speech recognition,” *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 5, pp. 360 –378, Sep. 1996.

[211] O. Ibe, *Markov Processes for Stochastic Modeling*. Elsevier, Sept. 2008.

[212] C. Grinstead and J. Snell, *Introduction to Probability: Second Revised Edition*. American Mathematical Society, 1997.

[213] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, pp. 107 – 117, 1998.

[214] Geodirectory, http://www.geodirectory.ie/, accessed Aug. 2012.

[215] A. Schlote, E. Crisostomi, S. Kirkland, and R. Shorten, "Traffic modelling framework for electric vehicles," *International Journal of Control*, vol. 85, no. 7, pp. 880–897, 2012.

[216] M. Ficek and L. Kencl, "Inter-Call Mobility model: A spatio-temporal refinement of Call Data Records using a Gaussian mixture model," in *INFOCOM, 2012 Proceedings IEEE*, 2012, pp. 469–477.

[217] L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macedo, B. Moelans, and A. Vaisman, "A model for enriching trajectories with semantic geographical information," in *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, ser. GIS '07. ACM, 2007, pp. 22:1–22:8.

[218] Dublinked, http://www.dublinked.ie/, accessed Jan 2013.

[219] R. G. Hollands, "Will the real smart city please stand up?" *City*, vol. 12, no. 3, pp. 303–320, 2008.

[220] A. Caragliu, C. Del Bo, P. Nijkamp *et al.*, *Smart cities in Europe*. Vrije Universiteit, Faculty of Economics and Business Administration, 2009.

[221] D. Washburn and U. Sindhu, "Helping CIOs Understand Smart City Initiatives," *Growth*, 2009.